

**COMPARING DIF DETECTION FOR MULTIDIMENSIONAL POLYTOMOUS
MODELS USING MULTI GROUP CONFIRMATORY FACTOR ANALYSIS AND THE
DIFFERENTIAL FUNCTIONING OF ITEMS AND TESTS**

by

Priya Kannan

M.S., Bangalore University, 2001

M.A., Minnesota State University, 2003

Submitted to the Graduate Faculty of the
School of Education in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This Dissertation was presented

by

Priya Kannan

It was defended on

April 4, 2011

and approved by

Dr. Kevin H. Kim, Associate Professor, Psychology in Education

Dr. Clement A. Stone, Professor, Psychology in Education

Dr. Suzanne Lane, Professor, Psychology in Education

Dr. Levent Kirisci, Professor, Pharmaceutical Sciences

Dissertation Advisor: Dr. Kevin H. Kim, Associate Professor, Psychology in Education

Copyright © by Priya Kannan

2011

**COMPARING DIF DETECTION FOR MULTIDIMENSIONAL
POLYTOMOUS MODELS USING MULTI GROUP CONFIRMATORY
FACTOR ANALYSIS AND THE DIFFERENTIAL FUNCTIONING OF ITEMS
AND TESTS**

Priya Kannan, PhD

University of Pittsburgh, 2011

This study evaluated the robustness of DIF detection for multidimensional polytomous items using two different estimation methods, MG-CFA and MGRM-DFIT. A simulation study across 960 study conditions was performed. The purpose of this study was to establish the Type-I error rate and Power of DIF detection for the MG-CFA and MGRM-DFIT estimation methods across the study conditions.

The MGRM-DFIT method consistently controlled Type-I error rate under alpha across all study conditions. Though the MGRM-DFIT method demonstrated high power in detecting DIF for the combined items, it had lower power in detecting DIF for each item individually. The MGRM-DFIT method had higher power of DIF detection when *impact* (true distributional differences) is in the opposite direction of manipulated DIF. Overall, compared to the non-DIF items, NCDIF values are larger, and CDIF values are smaller for the 4 DIF items. Across the replications and the study conditions, CDIF was not as consistent as NCDIF.

The MG-CFA method demonstrated slightly inflated Type-I error rate in a couple of study conditions (particularly in the presence of impact). However, the MG-CFA method demonstrated lower power across all study conditions. This could partly be explained by the low magnitude of DIF that was manipulated in the ' α/λ ' parameter in this study.

Parameter estimation for the MGRM, and the MGRM-DFIT method should be incorporated as part of commonly used software packages. In general, the MG-CFA method is recommended for DIF detection with multidimensional polytomous types of items, since it performs more consistently as a univariate test and as a multivariate test, and is easily available as part of several commonly used software packages.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	XIV
1.0 INTRODUCTION.....	1
1.1 PROBLEM STATEMENT	1
1.2 RESEARCH QUESTIONS.....	5
2.0 REVIEW OF LITERATURE	7
2.1 PERFORMANCE ASSESSMENTS	8
2.2 DIFFERENTIAL ITEM FUNCTIONING.....	10
2.3 THE MULTIDIMENSIONAL GRADED RESPONSE MODEL (MGRM) 14	
2.4 THE CONFIRMATORY FACTOR ANALYSIS (CFA) FRAMEWORK	18
2.5 RELATIONSHIP BETWEEN CFA AND IRT PARAMETERS	19
2.6 THE MULTI-GROUP CONFIRMATORY FACTOR ANALYSIS (MG-CFA) MODEL	20
2.6.1 Tests of Measurement Invariance	23
2.6.1.1 Omnibus test of Invariant Covariance matrices $\Sigma g = \Sigma g'$..	23
2.6.1.2 Configural Invariance $\Lambda form g = \Lambda form g'$	24
2.6.1.3 Metric Invariance $\Lambda g = \Lambda g'$	25

2.6.1.4	Invariant unique variance $\Theta\delta g = \Theta\delta g'$	28
2.6.2	Tests of Structural Invariance.....	28
2.6.2.1	Invariant factor variance $\Phi p p g = \Phi p p g'$	28
2.6.2.2	Invariant factor covariance $\Phi p p' g = \Phi p p' g'$	29
2.6.2.3	Invariant factor means $\kappa g = \kappa g'$	29
2.7	COMMONLY USED DIF DETECTION METHODS FOR DICHOTOMOUS AND POLYTOMOUS SCALES.....	31
2.7.1	Traditional DIF detection procedures	32
2.7.1.1	ANOVA and ANCOVA-based methods.	32
2.7.1.2	Delta-plot or Transformed Item Difficulty (TID) method. .	33
2.7.2	Nonparametric DIF detection procedures	34
2.7.2.1	Mantel-Haenszel (MH) procedure.....	34
2.7.2.2	Standardization method.	36
2.7.2.3	The non-parametric Poly-SIBTEST procedure.....	37
2.7.3	IRT-based DIF detection procedures for dichotomous and polytomous items.....	38
2.7.3.1	The LR procedure.....	40
2.8	DIFFERENTIAL FUNCTIONING OF ITEMS AND TESTS (DFIT) PROCEDURE.....	41
2.8.1	The DFIT method for dichotomous items	42
2.8.1.1	Comparison of the CDIF and NCDIF measures.....	47
2.8.2	The GRM-DFIT procedure	48
2.8.3	The Multidimensional DFIT procedure for dichotomous items ...	50

2.8.4	Multidimensional DFIT for polytomous items as an extension of GRM-DFIT	53
2.8.5	DFIT Significance Tests	57
2.9	PERFORMANCE OF THE DFIT METHODS AND COMPARISONS TO OTHER DICHOTOMOUS AND POLYTOMOUS DIF DETECTION PROCEDURES.....	58
2.10	CFA-BASED APPROACHES FOR DIF DETECTION AND SOME COMPARISON WITH IRT-BASED DIF DETECTION	66
3.0	METHOD	81
3.1	DESIGN	82
3.2	DATA GENERATION.....	91
3.3	DATA ANALYSIS.....	106
3.3.1	Outcome measures recorded.	106
3.3.2	Type-I error rate and Power.	107
3.3.3	Generalized Estimating Equations (GEE).	107
4.0	RESULTS	109
4.1.1	Within-Subject effects.	111
4.1.2	Between-Subject effects.....	124
4.1.3	CDIF and NCDIF.	127
5.0	DISCUSSION	130
5.1	SUMMARY OF MAJOR FINDINGS	131
5.1.1	Type-I error rates and empirical power for the MGRM-DFIT estimation method summarized across study conditions.	131

5.1.1.1	CDIF and NCDIF.....	133
5.1.2	Type-I error rates and empirical power for the MG-CFA estimation method summarized across study conditions.	133
5.1.3	Effect of Independent Variables.....	136
5.2	LIMITATIONS, FUTURE DIRECTIONS AND PRACTICAL IMPLICATIONS.....	137
APPENDIX A	142
	RESULTS FROM THE LOGISTIC REGRESSION	142
APPENDIX B	150
	SAS CODE FOR DATA GENERATION	150
APPENDIX C	159
	MPLUS CODES	159
APPENDIX D	166
	SAS CODE FOR READING-IN DATA FROM MPLUS.....	166
APPENDIX E	178
	SAS CODES FOR RUNNING THE MGRM-DFIT MACROS.....	178
APPENDIX F	205
	SAS CODE FOR COMPILING THE SIMULATION	205
BIBLIOGRAPHY	210

LIST OF TABLES

Table 3-1 Study Design	83
Table 3-2 Simulation Flow Chart	92
Table 4-1 Proportion of replications where DIF was detected for each Dependent Variable across the Sample size, Mean Difference, and Estimation method conditions.	113
Table 4-2 Proportion of replications where DIF was detected for each Dependent Variable across the Mean Difference and Estimation Method conditions.	116
Table 4-3 Proportion of replications where DIF was detected for each Dependent Variable across the Uniform DIF, Latent Mean Difference, and Estimation Method conditions.....	119
Table 4-4 Proportion of replications where DIF was detected for each Dependent Variable across the Uniform DIF and Estimation method conditions.....	120
Table 4-5 Proportion of replications where DIF was detected for each Dependent Variable in the control condition across the two Estimation methods.....	123
Table 4-6 Proportion of replications where DIF was detected for each Dependent Variable averaged across the study conditions for the two Estimation methods.....	123
Table 4-7 Proportion of replications where DIF was detected for each Dependent Variable across the Mean Difference and Sample size conditions.	125

Table 4-8 Proportion of replications where DIF was detected for each Dependent Variable across the Mean Difference conditions.	126
Table 4-9 Proportion of replications where DIF was detected for each Dependent Variable across the Uniform DIF conditions.	126
Table 4-10 CDIF and NCDIF values for the 14 items across study conditions	127

LIST OF FIGURES

Figure 3-1 Category Response Functions for a 5-point item for the <i>reference</i> and <i>focal group</i> when no DIF is introduced.	96
Figure 3-2 Category Response Functions for a 5-point item for the <i>reference</i> and <i>focal group</i> when no DIF is introduced in the 'a' parameter, and DIF of 0.5 is introduced in the highest 'b' parameter.....	97
Figure 3-3 Category Response Functions for a 5-point item for the <i>reference</i> and <i>focal group</i> when no DIF is introduced in the 'a' parameter, and DIF of 0.5 is introduced in the two highest 'b' parameter.....	100
Figure 3-4 Category Response Functions for a 5-point item for the reference and focal group when a DIF of 0.2 is introduced in both 'a' parameters, and DIF of 0.5 is introduced in the highest 'b' parameter.....	101
Figure 3-5 Category Response Functions for a 5-point item for the reference and focal group when a DIF of 0.2 is introduced in both 'a' parameters, and DIF of 0.5 is introduced in the two highest 'b' parameter.....	102
Figure 4-1 DIF detection for each DV across the Latent Mean Difference and Estimation method conditions when N=2000	115
Figure 4-2 DIF detection for each Dependent Variable across the Uniform DIF and Estimation method conditions	122

Figure 4-3 CDIF and NCDIF values for all 14 items averaged across replications 128

LIST OF ABBREVIATIONS

2PL - 2-parameter Logistic Model

2P-SRM - 2-parameter Sequential Response Model

ANOVA - Analysis of Variance

ANCOVA - Analysis of Covariance

CDIF - Compensatory DIF index

CFA - Confirmatory Factor Analysis

CRF - Category Response Function

CTT - Classical Test Theory

D - Difference in Scores

DIF - Differential Item Functioning

DFIT - Differential Functioning of Items and Tests

DTF - Differential Test Functioning

EPC - Expected Proportion Correct

ES - Expected Score

ES_F - Expected Score for the Focal Group

ES_R - Expected Score for the Reference Group

FA - Factor Analysis

FCAT - Florida Comprehensive Assessment Test

FG - Focal Group

FP - False Positives

GEE - Generalized Estimating Equations

GPCM - Generalized Partial Credit Model

GRM - Graded Response Model

GRM-DFIT - Graded Response Model based Differential Functioning of Items and Tests

GRM-LR - Graded Response Model based Likelihood Ratio Test

ICC - Item Characteristic Curves

IRT - Item Response Theory

IRT-LR - Item Response Theory based Likelihood Ratio Test

LR - Likelihood Ratio

M2PL - Multidimensional 2-parameter Logistic Model

MACS - Means and Covariance Structure Analysis

ME/I - Measurement Equivalence / Invariance

MI - Modification Index

MIRT - Multidimensional Item Response Theory

MG-CFA - Multi Group Confirmatory Factor Analysis

MGRM - Multidimensional Graded Response Model

MGRM-DFIT - Multidimensional Graded Response Model based Differential Functioning of
Items and Tests

M-H - Mantel-Haenszel Procedure

MSPAP - Maryland School Performance Assessment Program

NAEP - National Assessment of Educational Progress

NCDIF - Non-Compensatory DIF index

PSSA - Pennsylvania System of School Assessment

RG - Reference Group

RMSD - Root Mean Square Deviation

SAS - Statistical Analysis Software

SEM - Structural Equation Modeling

SSR - Sample Size Ratio

SRM - Sequential Response Model

TCC - Test Characteristic Curves

TID - Transformed Item Difficulty

TP - True Positives

WLSMV - Weighted Least Squares, Means and Variance adjusted

ACKNOWLEDGEMENTS

I would like to first thank my advisor, Dr. Kevin Kim for his constant guidance, every step of the way, throughout the six years of my Ph.D. I am extremely grateful to Kevin, for instilling my research interests in the field of ‘Structural Equation Modeling’, which helped model part of my Dissertation Research. I am thankful to Kevin, particularly for his never ending patience during my initial years in the program. I would also like to thank Kevin for going out of his way, all the time, to help all students. Kevin is one of the best advisors, and this formidable academic effort would have been impossible without his constant encouragement and support.

I am very grateful to Dr. Clement Stone, Dr. Suzanne Lane, and Dr. Levent Kirisci for being part of my committee and providing valuable suggestions for improving my work. I would like to render special thanks to the RM faculty members. I am grateful to Dr. Clement Stone for introducing me to the field of ‘Item Response Theory’, and instilling my interests in the ‘Graded Response Model’. I would like to thank Dr. Suzanne Lane for constantly believing in me. I am truly grateful to her for the variety of opportunities and experiences in the last six years as a result of her recommendation. Finally, I would like to thank Dr. Feifei Ye for her support, both as a teacher and as a friend. Thank you Feifei, for being a wonderful teacher, and also for all the sumptuous lunches and dinners I will never forget!

I am grateful for having had the opportunity to take courses under Dr. Lou Pingel, Dr. Carol Baker, Dr. Laura Hamilton, Dr. Lindsay Matsumura, Dr. Feifei Ye, Dr. Suzanne Lane, Dr. Clement Stone, and Dr. Kevin Kim. The valuable knowledge gained through the projects undertaken in these classes, and the ensuing conference presentations served as formative steps to my Dissertation research. I am also grateful to Dr. Jennifer Cartier and Dr. Ellice Forman for their support during my years of work on the NSF grant at DIL. I would like to especially thank Jen for supporting all my travel grants to conferences to present both my own research, and my collaborative work with DIL.

I would like to thank Ryan, Sean, Xiaowen, Laura, Lauren, Debra, Liquun, and Salma, for making my days in the program fun and exciting. I would like to also wish the current students all success in their endeavors, both during their Ph.D, and afterward. I would like to thank Leslie for all the fun times at DIL, for the late night meetings, for the spontaneous party streamers, for the girlscout cookies, for my send-off from Pittsburgh, and for being a wonderful friend! Finally, I would like to thank Barb for, first being a wonderful person and friend, and more importantly for managing all the administrative paperwork and arrangements smoothly from the beginning to the end.

I would like to thank my friends, Amrita, Aarti, and Shruthi for being my support system, and my extended family at Pittsburgh. I would like to thank Melanie, my roommate, for being a great friend. I would like to thank Roli for her hospitality during my first few years in Pittsburgh. I would also like to thank my other friends, Sonal, Becky, Debrup, Kaustubh, Akiko, Bob, Meg, Catherine and Geeta for making Pittsburgh my home for six years. Thank you for all the parties, potlucks, dinners, late nights at coffee shops, ballets, Broadway shows, barbeques, wine tastings, and all the other fun

events, which helped save my sanity during academic meltdowns ☺. I would especially like to thank Amrita for being there through everything!!!

I would like to thank my parents and my brother for being my constant support system, and my life-long cheer leaders! I would also like to thank my in laws (parents-in-law and sisters-in-law) for their support and encouragement. My parents' incessant belief in my abilities always proved to be a morale boost at times of distress. Appa and Amma, you have been the greatest parents! Thank you both for everything!!!

Finally, and most importantly, I would like to thank my wonderful husband, Mani. Mani's encouragement and support, every step of the way, was vital in helping me complete my Ph.D. I am indebted to him for all the sleepless nights that he shared with me, just so I could get my simulation to work correctly. During the last year, when I had to work on my research from New Jersey, Mani was my pseudo-advisor, my network, my support group, and my chauffeur to Pittsburgh for the Proposal and Defense meetings. Mani, thank you so much for all this, and everything else that I could never cease to list... Thank you for being there, and thank you for being you!!!

I dedicate this work to my husband and my parents, whose love and support guided me through this academic milestone.

1.0 INTRODUCTION

1.1 PROBLEM STATEMENT

Measurement processes are typically aimed at describing individuals and groups based on certain traits and characteristics that interest the researchers. In educational research and practice, assessment of students' knowledge and skills is pivotal to establishing mastery over the given field of study (e.g., mathematics, reading, science). Items and tests are developed in order to assess the achieved mastery in a field of study. The scores obtained from a test help determine the student's mastery of the subject area. But this is only to the extent that the test is a valid and reliable indicator of the underlying trait being measured. Psychological research often focuses on measuring constructs that cannot be directly observed (such as depression, anxiety, personality). Since these constructs cannot be directly measured, they are operationalized using a set of items. Again, these items provide us useful information about the individual's level on the construct provided that the items are valid indicators of the construct. In order to establish the validity of these items as indicators of the latent construct, measurement models (such as, item response theory, structural equation modeling) have been proposed.

However, an additional complication arises when one seeks to make group comparisons. Researchers in psychology have long been interested in comparing groups

and subgroups of various cultures on these constructs (e.g., Hofstede, 2001), and making some inferences on group similarities and differences on the trait of interest. In order to compare groups, however, one must ascertain that the numerical values being assigned are on the same measurement scale. In other words, one must be able to assume that any test/assessment has “measurement invariance” (Reise, Widaman, & Pugh, 1993; Vandenberg & Lance, 2000). Furthermore, one also needs to make sure that the items themselves are not differentially accessible to the sub-groups of interest. Within educational testing, by the 1960s, it was becoming apparent that there was a huge difference in the mean scores of children from Caucasian backgrounds, when compared to African-American and Latino children. Early studies on subgroup comparison were undertaken around the 1960s as the psychometric community’s response to public concern that the cognitive abilities assessed in these test items were outside the realms of the common experiences in minority cultures (Angoff, 1993). Such tests were deemed as unfair to the minority examinees since the items focused on skills and abilities that minority children had little opportunity to learn (Angoff, 1993).

Over time, Differential Item Functioning (DIF) analysis has been conducted in educational research with the primary goal of finding items that function differentially across groups, and possibly excluding those items from the final test. However, DIF analysis has focused primarily on dichotomous items (Camili & Congdon, 1999; Oshima, Raju, & Flowers, 1997; Raju, van der Linden, & Fleer, 1995), and rarely focused on multidimensional tests (Wu & Lei, 2009). But the truth of the fact is that, dimensionally complex tests measuring more than one latent trait have become the rule rather than the

exception in educational testing (McKinley & Reckase, 1983; Reckase, 1985; Reckase, 1987).

Furthermore, performance assessments have a significant advantage over multiple-choice questions in assessing the student's complex knowledge and skills (Lane & Stone, 2006). Consequently, performance assessments are used in large scale assessments and accountability programs. Performance assessments usually tend to cover a broader array of topics and each assessment typically tends to simultaneously assess multiple skills (Lane & Stone, 2006). Therefore, performance assessments are more likely to be multidimensional. These complex multidimensional-polytomous tests also have to be validated and examined for invariance in measurement properties across multiple subgroups.

Wu and Lei (2009) point out that the unique challenge of assessing DIF for multidimensional tests has rarely been investigated within educational research. Furthermore, assessing DIF for multidimensional-polytomous tests through traditional methods such as Item Response Theory (IRT) is typically challenging (Reckase, 1985, 1987) due to the assumptions of unidimensionality for most IRT models. The Multidimensional Graded Response Model (MGRM) has been proposed as an alternative to the GRM, and is said to handle multidimensionality more efficiently (DeAyala, 1994). However, a DIF assessment technique for the MGRM is yet to be proposed.

Among the IRT-based DIF detection methods, Raju's Differential Functioning of Items and Tests (DFIT) method is probably the most frequently used technique (Flowers, Oshima, & Raju, 1999; Oshima, Raju, & Flowers, 1997; Raju, van der Linden, & Fleer, 1995). This method, originally proposed for unidimensional dichotomous items (Raju,

van der Linden, & Fleer, 1995), was later extended to unidimensional polytomous items (Flowers, Oshima, & Raju, 1999), and to multidimensional dichotomous items (Oshima, Raju, & Flowers, 1997). However, the DFIT method has not yet been extended for multidimensional polytomous items and the MGRM. Oshima, Raju, and Flowers (1997) proposed a DFIT method for multidimensional-dichotomous items, and assessed the robustness of this method in assessing DIF for intentionally multidimensional-dichotomous tests. They proposed both compensatory (CDIF) and non-compensatory (NCDIF) measures for assessing DIF when dichotomous items are multidimensional. Both their CDIF and NCDIF measures were effective in controlling Type-I error rates in detecting DIF. However, both their CDIF and NCDIF measures had low power ($\leq .75$) in detecting DIF (Oshima, Raju, & Flowers, 1997).

Recently, there has also been a growing interest in applying Structural Equation Modeling (SEM)'s Multi Group-Confirmatory Factor Analysis (MG-CFA) approach in DIF investigations (Chan, 2000; Gonzalez-Roma, Hernandez, & Gomez-Benito, 2006; Wu & Lei, 2009). SEM has been found to have comparatively more flexibility in handling multiple latent constructs (Kannan & Kim, 2009; Kannan & Ye, 2008; Raju, Laffitte, & Byrne, 2002). Gonzalez-Roma, Hernandez, and Gomez-Benito (2006) assessed the power and Type-I error rate for detecting DIF for a unidimensional Graded Response Model (GRM). They found that the MG-CFA approach had acceptable power ($\geq .70$) in detecting DIF even for conditions with small samples and medium levels of DIF. Furthermore, they found that the power in DIF detection under the MG-CFA approach increased as sample size and DIF magnitude increased (Gonzalez-Roma, Hernandez, & Gomez-Benito, 2006). Finally, they found that the Type-I error rate was

consistently controlled ($<.10$ in all cases) under the MG-CFA model. Wu and Lei (2009) used the MG-CFA approach to detect DIF for multidimensional dichotomous models. They compared the power and Type-I error of the MG-CFA model in detecting DIF when the model was misspecified as unidimensional to when it was correctly specified as two-dimensional. They found that Type-I error was significantly reduced and power significantly increased when the two-factor MG-CFA model was used (Wu & Lei, 2009).

1.2 RESEARCH QUESTIONS

The burgeoning interest in assessing DIF using the MG-CFA approach is promising for investigators who work on DIF detection, but are encountered with multidimensional models. However, no study has investigated the performance of MG-CFA models in assessing DIF for multidimensional polytomous models. Furthermore, Raju's DFIT method that exists for multidimensional dichotomous models has not been extended to multidimensional-polytomous models to assess the effectiveness of a MGRM-based approach to DIF detection. Polytomous models pose different challenges to researchers when compared to dichotomous models. Nevertheless, with the increasing interest in performance-based assessments, polytomous models are largely being used. Therefore, assessing and comparing the robustness of MG-CFA based and MIRT based approaches for DIF detection in multidimensional polytomous models is a very important and relevant research question that needs to be addressed.

This study, therefore, had two main purposes: (1) to extend Raju's DFIT technique and propose an IRT based DIF detection method for multidimensional polytomous items

and tests; and (2) to compare the performance of MG-CFA based approaches to the MGRM-based DFIT approach in DIF detection for multidimensional polytomous tests.

Specifically, the following research questions were addressed:

- (1) What is the *Type-I error rate* and *Power* of the MGRM-based DFIT approach in identifying DIF items?
- (2) What is the *Type-I error rate* and *Power* of the MG-CFA-based approach in identifying DIF items?
- (3) Is the pattern of difference on *Type-I error rate* and *Power* between the MGRM-DFIT and the MG-CFA different among the levels of the Independent variables considered (*sample size, sample size ratio, type of DIF, DIF direction, and differences in latent distribution*)?

2.0 REVIEW OF LITERATURE

Within the context of classical test theory (CTT; Crocker & Algina, 1986), an abstract hypothetical construct, ' τ ' or ' ξ ', is measured using an observed variable ' X '. It is also assumed that the measurement systems are imperfect, and some proportion of the variation in ' X ' are typically attributed to systematic and unsystematic measurement error, ' E ' or ' δ '. In comparing different groups on the variable of interest, several important assumptions are made. The presence or absence of group differences is assumed to have substantive implications. Furthermore, the measure is assumed to comprise of multiple manifest indicators (or items), which are combined additively to operationalize the underlying construct. Finally, it is assumed that the psychometric soundness (reliability and validity) of the measure can be demonstrated (Vandenburg & Lance, 2000). The important research questions that underlie group comparisons, especially for complex multidimensional polytomous models, cannot be directly addressed within the traditional framework of CTT (Vandenburg & Lance, 2000). However, with recent advances in analytic tools such as item response theory (IRT) and confirmatory factor analysis (CFA), these hypotheses are now more testable.

In the following sections a detailed literature review is presented. First, performance assessments, which justify the need to develop tests for multidimensional polytomous models are briefly introduced. This is followed by an introduction of Differential Item

Functioning (DIF) and its importance in educational testing. In the next sections, the Multidimensional Graded Response Model (MGRM) and the CFA model are introduced, and the equivalence between the MGRM and the CFA models are established. This is followed by a review of the MG-CFA-based measurement invariance approach to testing invariance of item properties across subgroups.

Next, some of the most commonly used, parametric and non-parametric, DIF detection approaches are reviewed. Specific attention is given here to some of the IRT-based DIF detection methods for dichotomous and polytomous models. The DFIT family of measures (Flowers, Oshima, & Raju, 1999; Oshima, Raju, & Flowers, 1997; Raju, van der Linden, & Fleer, 1995), for the unidimensional dichotomous and polytomous tests are then presented. Subsequently, the logical extension of Raju's DFIT method for multidimensional polytomous items from the previously existing equations for multidimensional dichotomous items (Oshima, Raju & Flowers, 1997) and unidimensional polytomous items (Flowers, Oshima & Raju, 1999) are presented. Finally, relevant literature assessing the performance of the DFIT method and the MG-CFA method in detecting DIF for unidimensional and multidimensional types of items are presented.

2.1 PERFORMANCE ASSESSMENTS

Performance assessments typically require students to perform tasks, as opposed to responding to a stimulus based on a provided selection of responses. Students are expected to perform in an emulated context and demonstrate their knowledge and skills

as applied to the task at hand (AERA, APA, & NCME, 1999). Tasks such as experiments, essays, portfolios, and extended-constructed response questions are considered as performance assessments. These performance tasks are said to emulate realistic applications that students will encounter in their academic or professional lives, and are therefore very '*meaningful*' for the students (Lane & Stone, 2006). Therefore, these assessments allow for a direct alignment between assessment and instructional activities. Performance assessments have influenced curriculum and instructional changes by encouraging teachers to broaden the focus of their teaching to include activities that elicit student reasoning and problem solving (Lane & Stone, 2006).

Many school districts and state and national assessments have incorporated performance assessments. The Advanced Placement (AP) exams used to determine high-school student proficiency in college courses consist of constructed response items. The National Assessment of Educational Progress (NAEP) which is a national assessment of student knowledge in a given subject also includes performance assessment items. Certain state assessment programs (e.g., Maryland) have used entirely performance based assessments in assessing their students. Furthermore, at the classroom level, performance assessments are used by teachers to diagnose student strengths and weaknesses in subject matter (Lane & Stone, 2006).

However, these performance assessments are typically assessed using an explicit rubric, and the student responses are scored on a scale that reflects several levels of performance (Lane & Stone, 2006). Therefore, these responses typically tend to be polytomous in nature. Furthermore, performance assessments usually tend to cover a broad array of topics. Each assessment typically tends to simultaneously assess multiple

skills or predictive factors (Lane & Stone, 2006). Therefore, such assessments are more likely to be multidimensional. Therefore, a multidimensional polytomous measurement model is imperative to assessing performance-based items.

2.2 DIFFERENTIAL ITEM FUNCTIONING

Angoff (1993) traces the roots of Differential Item Functioning (DIF) analyses to the “cultural difference” studies that were first undertaken in the early 1960s. It was becoming apparent that there was a huge difference in the mean scores of Caucasian children when compared to African-American and Latino children. These early studies (in the 1960s) were undertaken as the psychometric community’s response to public concern that the cognitive abilities assessed in these test items were outside the realms of the common experiences in minority cultures (Angoff, 1993). These tests were deemed as unfair to the minority examinees since the items focused on skills and abilities that minority children had little opportunity to learn (Angoff, 1993).

With the Civil Rights and Feminist movements, the sensitivity to racial (and gender) issues have resulted in increased interest in assessing bias in tests, and creating educational curriculum that is fair to all the subgroups concerned (Cole, 1993). The main goal of these early studies was to find items on a test that are ‘biased’ toward minority examinees, and to remove them from the test (Angoff, 1993; Cole, 1993; Walker & Beretvas, 2001). In current practice, identifying items which exhibit bias is only a preliminary step in assessing item and test bias. The ultimate rationale is that the removal

of these biased items will improve the overall validity of the test, and will eventually result in a test that is fair to all examinees (Camili & Congdon, 1999).

The highly politicized environment in which item bias was being examined resulted in several controversies around the usage of the term ‘bias’, due to the semantic conflict in the social and statistical implications of the term (Angoff, 1993; Zumbo, 2007). The expression ‘Differential Item Functioning’ (DIF) was consequently introduced to refer to the item that displays different statistical properties for different subgroups, after controlling for the abilities of the subgroups in consideration (Angoff, 1993; Bolt, 2002; Zumbo, 2007). DIF is considered a relative term since it is always used when comparing one group of examinees to another on a given item (Holland & Wainer, 1993).

When two groups are measured to have the same amount of the underlying trait, but perform differently on any item, then DIF is said to occur for the given item (Bolt, 2002). DIF is therefore a statistical term used to refer to a situation where persons from one group have a higher probability of getting an item correct, when compared with persons, of equal ability, from another group (Zumbo, 2007). Or in other words, for two examinees of comparable ability, a given test item is said to demonstrate DIF, if the probability of a correct response on that item is associated with group membership (Camili, 1992; Camili & Congdon, 1999).

It can therefore be assumed that these (DIF) items possibly measure one (or more) irrelevant constructs, in addition to the target trait being measured by the test (Camili, 1992; Walker & Beretvas, 2001). These additional nuisance-constructs account for the difference in item performance for two examinees of otherwise equal ability. Therefore, the introduction of the term DIF allowed one to distinguish ‘*item-impact*’ from ‘*item-*

bias'. Item impact is said to exist when there are true differences between groups in the underlying ability of interest, whereas DIF or item bias exists when some characteristic of the test item, not relevant to the measured underlying ability, causes the groups to differ in their performance (Camili, 1992; Walker & Beretvas, 2001; Zumbo, 2007).

In DIF analysis, it is customary to refer to the examinee group of interest, typically the minority group, as the *focal group*. The group to which their performance is being compared is referred to as the *reference group* (Bolt, 2002; Holland & Wainer, 1993; Raju, van der Linden, & Fler, 1995). Typically, in any study, there could be multiple *focal/reference* pairs of groups for which DIF comparisons can be made. In general, DIF detection methods are involved in testing null and alternate hypotheses of the following form: “ H_0 : The item functions equally for the *reference* and *focal* groups (no DIF)”, and “ H_A : The item functions unequally for the *reference* and *focal* groups (DIF)” (Bolt, 2002, pp. 115).

Two types of DIF, *uniform* and *non-uniform* DIFs are commonly investigated in DIF research (Gonzalez-Roma, Hernandez, & Gomez-Benito, 2006; Wu & Lei, 2009; Zumbo, 2007). In the initial conception of *item bias*, and ANOVA-based DIF comparisons, the main-effects of group differences were referred to as *uniform* DIF, and the interaction between group and ability was referred to as *non-uniform* DIF (Zumbo, 2007). However, within the IRT framework, *uniform* and *non-uniform* DIFs are used with *references* to differences in the intercept and discrimination parameters across groups. When the groups differ only in the intercept or item difficulty parameters, then this is termed as *uniform* DIF (Gonzalez-Roma, Hernandez, & Gomez-Benito, 2006). However, when

there are differences in the item discrimination parameters, this is referred to as *non-uniform* DIF (Gonzalez-Roma, Hernandez, & Gomez-Benito, 2006).

Zumbo (2007) summarized the major trends in DIF research, and classifies DIF research into three generations, based on where DIF analysis originally started, where the current state of research lies, and where the research should be headed. During what Zumbo calls the first generation of DIF analysis, *item bias* was still the commonly used term to refer to DIF. Furthermore, attention was mainly paid to two groups of examinees and dichotomous items during the first generation of DIF (Zumbo, 2007). During this time, ANOVA (Cardall & Coffman, 1964) and ANCOVA (Angoff & Sharon, 1974) procedures were mainly used to test interaction terms for differences in a subgroup performance (Angoff, 1993). The *delta-plot* method or *transformed item difficulty* (TID) method (Angoff, 1972) also became popular around this time.

The transition to the second generation in DIF analysis was marked by the widespread acceptance of the term DIF rather than *item bias* (Zumbo, 2007). This generation of DIF research was marked mainly by separation of the intentional impact from unintentional bias. Furthermore, this generation of DIF research was marked by increased research focus on developing new and sophisticated statistical methods for identifying items with significant DIF (Zumbo, 2007). One of the most prominently used techniques, the Mantel-Haenszel method (Holland & Thayer, 1988), involved the use of contingency tables (Zumbo, 2007). The logistic regression (LR) based approaches (Thissen, Steinberg, & Gerrard, 1986), and other IRT-based approaches (Flowers, Oshima, & Raju, 1999; Oshima, Raju, & Flowers, 1997; Raju, van der Linden, & Fler, 1995) also began to gain prominence around this time (Zumbo, 2007).

Zumbo (2007) points out that the third generation of DIF research, (where the current interest lies, and where the field should be heading), is marked by a focus on multidimensional models, and applying SEM frameworks to DIF assessment. He also points out that research is and should be focused on multiple indicators, mixture modeling, and considering situational factors and external variables that impact test performance. He claims that DIF should be considered as an empirical method for investigating lack of invariance, model-data fit and appropriateness for all measurement frameworks (Zumbo, 2007). It can therefore be seen that the current state of DIF research lies in the area of DIF detection for multidimensional models. As it has been pointed out before, a huge gap exists in the literature when it comes to DIF detection methods for multidimensional polytomous models. In order to review DIF methods for multidimensional polytomous tests, some prominent multidimensional polytomous models, namely the Multidimensional Graded Response Model (MGRM) and the Confirmatory Factor Analysis (CFA) model, are reviewed here.

2.3 THE MULTIDIMENSIONAL GRADED RESPONSE MODEL (MGRM)

Item Response Theory (IRT) constitutes a collection of models that relate an examinee's item responses to his/her latent ability. This family of models is most prominently used in educational and psychological testing today (Embretson & Reise, 2000). These models make some strong assumptions about dimensionality and local independence of data, and when these assumptions are violated, the inferences made from these models become

questionable. When it comes to polytomously scored data, several unidimensional IRT models are available for scoring student responses and estimating their latent ability. Some commonly used polytomous models are: (1) the partial credit model (Masters, 1982); (2) the generalized partial credit model (Muraki, 1992); (3) the Graded Response Model (GRM; Samejima, 1969; 1972); (4) the modified GRM (Muraki, 1990); (5) the rating scale model (Andrich, 1978); and (6) the nominal response model (Bock, 1972). Due to its ease in analyzing items with multiple score categories, the GRM tends to be most frequently used for polytomous data, especially estimating item parameters for *ordered categorical responses* such as a Likert-type rating scale and performance assessments.

The GRM (Samejima, 1969; 1972) is widely used for unidimensional polytomous variables. The GRM is a generalization of the 2-parameter (2P) logistic model, where each item has ‘ k ’ ordered categorical response options. Each scale item (i) is described by one item slope parameter (α_i) and $k_i - 1$ between category “threshold” parameters (β_{ij}), $j = 1, 2, \dots, k_i - 1$. Samejima (1969; 1972) proposed a two-stage process to obtain the probability of predicting a given examinee’s score level. In the first stage, a between-category boundary score is estimated for each examinee. This is done by estimating the probability $P_{ij}^*(\theta)$ that an examinee receives a category score j ($j = 1, 2, \dots, k_i - 1$) or higher. The general form of the unidimensional GRM is given by

$$P_{ij}^*(\theta) = \frac{\exp(D\alpha_i(\theta - \beta_{ij}))}{1 + \exp(D\alpha_i(\theta - \beta_{ij}))} \quad (1)$$

where D is a scaling constant of 1.7.

The β_{ij} parameters represent the ability level (θ) necessary to have a 50% chance of responding in a category above the j^{th} between-category boundary. In the second stage, the probability of responding in a particular category conditional on θ is estimated. This is obtained easily by subtraction as follows:

$$P_{ij}(\theta) = P_{ij}^*(\theta) - P_{i(j+1)}^*(\theta) \quad (2)$$

where by definition, the probability of responding in or above the lowest category is $P_{i0}^*(\theta) = 1.0$, and the probability of responding above the highest category $P_{ik_i}^*(\theta) = 0.0$ (Embretson & Reise, 2000).

However, most performance assessments tend to assess multiple traits or skills. Furthermore, the complexity of the performance task also often tends to contribute to the multidimensionality of these assessments (Lane & Stone, 2006). For example, a mathematics performance item is likely to measure both mathematical problem solving and mathematical communication skills (Walker & Beretvas, 2001). It must be pointed out that both these traits are intentionally measured in these assessments, and is different from unintentional construct irrelevant factors that might result in DIF. Additionally, performance assessments are typically developed where a single item assesses multiple content areas. It can be seen that, multidimensionality in item responses for performance assessments is more than likely to be expected. Therefore, a multidimensional form of the GRM is important to consider.

For the multidimensional form of the GRM, assume that a set of H latent traits determine test performance, then the ability level for person 's' on the H latent traits are represented by a vector of values $\Theta_s = (\theta_{s1}, \dots, \theta_{sH})^T$. These values are considered to represent a random sample drawn from a population with a multivariate normal density

function, $g(\theta_s) \sim N(\mu, \Sigma)$, where μ and Σ represent the mean and the covariance matrix of Θ_s . The general form for the MGRM is given by:

$$P_{ij}^*(\Theta) = \frac{\exp[D \sum_{h=1}^H \alpha_{ih}(\theta_h - \beta_{ij})]}{1 + \exp[D \sum_{h=1}^H \alpha_{ih}(\theta_h - \beta_{ij})]} \quad (3)$$

where θ_h is the latent trait on dimension h ($h = 1, \dots, H$ dimensions); α_{ih} is the discrimination parameter for item i on dimension h ; β_{ij} is the threshold parameter for responding in category j for item i ; and the summation is over all the H dimensions. Within the multidimensional framework, $P_{ij}^*(\Theta)$ is the probability of a randomly selected examinee with latent traits Θ_s responding in category j or higher, for any given item i (De Ayala, 1994).

De Ayala (1994) assessed the parameter recovery of the MGRM for data that were generated from one-, two-, and three-dimensions. Regardless of the dimensionality of the data, he found that the β_{ij} parameters were estimated accurately for all models. However, as the number of factors increased, the overall correlation between the estimated and true discrimination parameters tended to fall. The estimated discrimination parameters were more strongly influenced by the mean of the discrimination parameter ($\bar{\alpha}$) value, than the value for the respective true discrimination parameter that was simulated (De Ayala, 1994).

2.4 THE CONFIRMATORY FACTOR ANALYSIS (CFA) FRAMEWORK

In a Factor Analysis (FA) framework, the model for ordinal responses takes the form:

$$x_i^* = \lambda_{i1}\xi_1 + \lambda_{i2}\xi_2 + \cdots + \lambda_{iH}\xi_H + \delta_i, i = 1, 2, \dots, p, \quad (4)$$

where x_i^* is an unobserved continuous response variable underlying the observed polytomous item x_i , ξ_h is the latent factor score, λ_{ih} is the factor loading of item i on factor h , it can also be seen as the coefficient representing the regression of x_i^* on ξ_h , and δ_i is an error term representing a specific factor and measurement error. There are a total of H factors. The observed discrete variable x_i is obtained by comparing the underlying variable x_i^* with threshold values τ_{ij} ,

$$x_i = j, \text{ if } \tau_{ij} < x_i^* < \tau_{i(j+1)}, j = 1, \dots, k_i - 1 \quad (5)$$

where k_i is the number of ordered-categorical responses for item i and $\tau_{i0} = -\infty, \tau_{ik_i} = +\infty$.

Within the FA framework, a complex structure with two factors would entail that the items are loaded on both factors and thus have two λ_i s per item, and the model becomes

$$x_i^* = \lambda_{i1}\xi_1 + \lambda_{i2}\xi_2 + \delta_i, i = 1, 2, \dots, p. \quad (6)$$

However, for the items that load only on one factor, the λ_{iH} term for the other factor resolves to zero. It should be noted that only a 2-factor model will be discussed in this paper. Rewriting Eq. (4) into matrix form:

$$x^* = \Lambda\xi + \delta, \quad (7)$$

where \mathbf{x}^* is a $(p \times 1)$ vector of items, Λ is a $(p \times H)$ matrix of loadings of the p measured variables on the H latent variables, ξ is a $(H \times 1)$ vector of factor scores, and δ is a $(p \times 1)$ vector of measurement residuals. Assuming that $E(\xi, \delta) = 0$, the covariance of items are:

$$\Sigma = \Lambda\Phi\Lambda' + \Theta_\delta \quad (8)$$

where Σ is the $(p \times p)$ population covariance matrix among the measured variables in Eq. (7), Φ is a $(H \times H)$ matrix of covariance among the latent variables, and Θ_δ is a diagonal matrix of unique variances (Reise, Widaman, & Pugh, 1993; Vandenberg & Lance, 2000).

2.5 RELATIONSHIP BETWEEN CFA AND IRT PARAMETERS

While the factor analysis parameters themselves do not correspond directly to the IRT item parameters, it is possible to transform the factor loadings λ and threshold values τ to obtain the item parameter estimates for the within-item multidimensional structure as follows. For a multidimensional model with two latent dimensions, the α_{ih} parameter in the MGRM can be expressed in terms of the factor loadings as

$$\alpha_{ih} = \frac{(D)\lambda_{ih}}{\sqrt{1 - (\sum_{h=1}^H \lambda_{ih}^2 + 2\lambda_{i1}\phi\lambda_{i2})}} \quad (9)$$

where, λ_{i1} contains the factor loading for i^{th} item on ξ_1 and λ_{i2} contains the factor loading for i^{th} item on ξ_2 ; and ϕ is the correlation between ξ_1 and ξ_2 (McLeod, Swygert, & Thissen, 2001; Swygert, McLeod, & Thissen, 2001; Takane & de Leeuw, 1987). Each complex item with two latent dimensions would have two α_{ih} or λ_{ih} slope parameters.

Further, for the 2-dimensional model, the item-category threshold parameters β_{ij} can be expressed in terms of CFA parameters as

$$\beta_{ij} = \frac{\tau_{ij}}{\sqrt{1 - (\sum_{h=1}^H \lambda_{ih}^2 + 2\lambda_{i1}\phi\lambda_{i2})}} \quad (10)$$

Kannan and Kim (2009) have shown that CFA with weighted least squares-means and variance adjustments (WLSMV) estimation method accurately estimates item parameters for the MGRM with low RMSD and bias for a variety of sample size, scale-point, correlation, and complex loading conditions. They found that the WLSMV estimation method produced consistently smaller standard errors, and took comparatively less time in estimating the item parameters, when compared to a Maximum Likelihood based estimation method. This advantage of the WLSMV method was especially apparent in conditions where the correlation between the latent dimensions was high (i.e., $\rho = .50$), and the complex loading on the secondary factor was high ($\lambda = .30$).

2.6 THE MULTI-GROUP CONFIRMATORY FACTOR ANALYSIS (MG-CFA) MODEL

In order to be able to interpret the observed group-mean differences unambiguously, it is important to establish between-group equivalence of the underlying measurement model. When group comparisons are made, one must ascertain that the numerical values being assigned are on the same measurement scale. In other words, it is important to establish the extent to which the measurement properties of the manifest variable are comparable or generalizable for each group ‘g’. Therefore, one must be able to assume that the

test/assessment has “Measurement Equivalence/Invariance (ME/I)” across groups (Reise, Widaman, & Pugh, 1993; Vandenberg & Lance, 2000). When the numerical values assigned to the trait scores are not comparable across groups, then the measured differences between groups might be artificial (Reise, Widaman, & Pugh, 1993; Vandenberg & Lance, 2000).

Therefore, to make group comparisons, Eq. (7) and (8) may be modified to denote group membership, giving:

$$x^g = \Lambda^g \xi^g + \delta^g \quad (11)$$

and

$$\Sigma^g = \Lambda^g \Phi \Lambda^{g'} + \Theta_{\delta}^g \quad (12)$$

where the superscript ‘g’ is added to Eq. (7) and (8) in order to indicate group membership.

The following assumptions about ME/I are typically invoked when testing hypotheses about group similarities and differences (Horn & McArdle, 1992; Vandenburg & Lance, 2000):

- The underlying theoretical latent variable (ξ^g) is conceptually equivalent in each group,
- The associations (λ^g) between ‘ x^g ’ and ‘ ξ^g ’ are equivalent across groups, and,
- The ‘ x^g ’ are influenced to the same degree by the same unique factors (δ^g) across groups.

In order to make cross-group comparisons, it is important to establish that the measurement operations are invariant across the groups being compared, and satisfy the above assumptions. If these assumptions are violated, then the conclusions drawn from group comparisons become debatable. Furthermore, the reliability of the scores and the validity of the score inferences also become questionable (Horn & McArdle, 1992; Vandenburg & Lance, 2000). Therefore, in order to satisfy the above assumptions, some

hypotheses about ME/I across groups must be tested. The following *testable hypotheses* regarding to ME/I are implied by Eq. (11) and (12):

1. $\xi^g = \xi^{g'}$, that is, the set of p items evokes the same conceptual framework in defining the latent construct (ξ) in each comparison group 'g'.
2. $\Lambda^g = \Lambda^{g'}$, that is, the factor loadings are same across groups.
3. $\tau^g = \tau^{g'}$, that is, the category thresholds are invariant across groups,
4. that the CFA model holds equivalently and assumes a common form across groups.
5. $\Theta_\delta^g = \Theta_\delta^{g'}$, that is, unique variances are invariant across groups.
6. $\Phi^g = \Phi^{g'}$, that is, variance and covariance of the latent variables are invariant across groups

Despite being readily testable, these aspects of ME/I are rarely evaluated in practice (Vandenburg & Lance, 2000). Furthermore, there is a lack of consensus on when the various tests of ME/I should be undertaken, and the order in which these tests should be undertaken (Cheung & Rensvold, 2002; Reise, Widaman, & Pugh, 1993; Vandenburg & Lance, 2000). Based on a meta-analysis of multiple statistical and applied papers, Vandenburg and Lance (2000) proposed eight ME/I tests that are most frequently undertaken in the literature. In most propositions of MG-CFA literature, there is general agreement that an omnibus test of the equality of covariance matrices should be undertaken first (Bagozzi & Edwards, 1998; Jöreskog, 1971; Vandenburg & Lance, 2000), and that this should be followed by tests of metric and scalar invariance.

Byrne, Shavelson, and Muthen (1989) proposed the distinction between tests of measurement invariance (tests that concern the relationship between the measured variables and the latent constructs), and tests of structural invariance (tests that concern the latent variables themselves). The omnibus test of equality of covariance matrices, and tests that assess metric and threshold invariance are typically considered tests of

measurement invariance. Factor means, variances, and covariances are usually examined as tests of structural invariance (Byrne, Shavelson, & Muthen, 1989). A detailed description of the seven most frequently used ME/I tests are as follows.

2.6.1 Tests of Measurement Invariance

2.6.1.1 Omnibus test of Invariant Covariance matrices $\Sigma^g = \Sigma^{g'}$.

The sample covariance matrices $S^g = S^{g'}$ are typically compared in applications of MG-CFA. The tenability of the null hypothesis ($\Sigma^g = \Sigma^{g'}$) is evaluated using a χ^2 statistic and other goodness-of-fit measures across multiple samples (Bollen & Long, 1993). No further tests are warranted when this null hypothesis holds. Failure to reject this null hypothesis is held as proof for the overall measurement equivalence of the multiple groups that are being compared (Vandenburg & Lance, 2000). Though the rejection of the null hypothesis for this omnibus test is indicative of some form of nonequivalence between the groups (Schmitt, 1982), it is uninformative otherwise. Specifically, it does not necessarily point out the particular source of nonequivalence that exists between the groups. Therefore, if the null hypothesis that $\Sigma^g = \Sigma^{g'}$ is rejected, then further ME/I testing, where a series of more restrictive hypotheses are tested, is warranted (Byrne, Shavelson, & Muthen, 1989). Finally, Vandenburg and Lance (2000), in their meta-analysis, found that although more than 62% of the statistical papers they reviewed

actually recommend conducting this omnibus test first, less than 20% of the applied studies actually used such a test (e.g., Cheung & Rensvold, 2002).

2.6.1.2 Configural Invariance $\Lambda_{form}^g = \Lambda_{form}^{g'}$.

Configural invariance is a test of the null hypothesis that the apriori pattern of fixed and free parameter loadings imposed on the items is equivalent across groups. It is also referred to as “Weak Factorial Invariance” (Horn & McArdle, 1992). It is a test of the overall fit that compares the form of the Λ matrix in both (or multiple) groups, and compares if the different groups similarly conceptualize the latent constructs. That is, the groups should have associated the same subsets of items with the same factors (Cheung & Rensvold, 2002). The factor structure (assuming that it is a reasonable representation of the underlying conceptual frame of *reference*) should also be comparable between groups (Vandenburg & Lance, 2000). Failure to reject the configural invariance null hypothesis that $\Lambda_{form}^g = \Lambda_{form}^{g'}$ has two implications: 1) Either that the respondents were using the same conceptual frame of *reference*, and therefore maybe compared; or 2) That further ME/I testing may be undertaken in order to ensure the comparability of the groups (Cheung & Rensvold, 2002; Vandenburg & Lance, 2000).

However, if the configural invariance null hypothesis is rejected, neither are the groups comparable, nor is additional ME/I testing warranted (Vandenburg & Lance, 2000). Configural invariance may fail, for example, when participants from different cultures (subgroups) attach different meanings and conceptual frames of *reference* to the constructs (Cheung & Rensvold, 2002). In addition, however, the null hypothesis of

$\Lambda_{form}^g = \Lambda_{form}^{g'}$ may also be rejected due to a host of other problems that include, but are not limited to, data collection problems, translation and back-translation errors, survey administration and instructional errors, to name a few (Cheung & Rensvold, 2002). If the underlying construct is not invariant between the groups, then comparing the groups on their performance is not very meaningful. Furthermore, it does not make any sense to occupy oneself in additional ME/I testing, and comparing if the items are calibrated similarly, when the underlying constructs are not comparable between the groups. Therefore, configural invariance must be established in order for the successive ME/I tests to be meaningful (Vandenburg & Lance, 2000).

2.6.1.3 Metric Invariance $\Lambda^g = \Lambda^{g'}$.

Metric invariance is a test of the null hypothesis that the factor loading parameters are equivalent across groups (Cheung & Rensvold, 2002; Vandenburg & Lance, 2000). It is a test of the strength of the relationship between items and their underlying construct, to see if the constructs are manifested in the same way across groups. The test of the null hypothesis $\Lambda^g = \Lambda^{g'}$ tests the equality of the scaling units across the groups (Schmitt, 1982; Vandenburg & Lance, 2000). Full-metric invariance is also referred to as “Construct-level Metric Invariance” (Cheung & Rensvold, 2002).

Data obtained from different populations may demonstrate conceptual agreement in terms of the type and number of underlying constructs, and the items associated with each construct (configural invariance). Still, the strength of the relationships between specific scale items and the underlying constructs may vary across groups (Cheung &

Rensvold, 2002). The test of full-metric invariance is achieved by constraining the factor loadings (λ_{ih}) of all ‘like’ items to be equal across groups. It is assumed that the model holds exactly for both (or all) groups in question, and that $\Lambda^g = \Lambda^{g'}$ holds unconditionally. Therefore, metric invariance is considered a stronger test than configural invariance (Horn & McArdle, 1992), and is important as a prerequisite for meaningful cross-group comparisons (Bollen, 1989).

Partial Metric Invariance.

The recommended course of action when the null hypothesis of $\Lambda^g = \Lambda^{g'}$ is rejected, is ambiguous in the literature (Reise, Widaman, & Pugh, 1993; Vandenberg & Lance, 2000). Some authors (e.g., Bollen, 1989) are of the opinion that the rejection of the null hypothesis for full metric invariance should preclude any further ME/I testing, much like when the null hypothesis of configural invariance is rejected. However, other authors (e.g., Byrne, Shavelson, & Muthen, 1989) have argued that “*partial metric invariance*” should be tested for when full-metric invariance does not hold. “Partial metric invariance” occurs if some, but not all, of the non-fixed values in Λ are invariant across groups (Byrne, Shavelson, & Muthen, 1989).

Cheung and Rensvold (2002) refer to partial invariance as “item-level metric invariance” or “factor loading invariance”. They suggest that a series of item-level metric invariance tests should be undertaken, if the “construct-level metric invariance” does not hold. This also enables one to locate the items responsible for the overall non-invariance of the factor loading matrix (Cheung & Rensvold, 2002). Item-level tests are enabled by most software programs, wherein Modification Indices (MIs) are computed for each fixed parameter (with 1 degree of freedom). These MIs indicate how much the overall χ^2 value

would change if a single constraint were added/removed (Reise, Widaman, & Pugh, 1993). This makes it possible for researchers to search for a subset of invariant items, for which the factor loadings do not vary across groups (Cheung & Rensvold, 2002; Reise, Widaman, & Pugh, 1993). In other words, the items that are not invariant across groups would be considered as items demonstrating DIF.

Testing for “partial metric invariance” serves as a control for measurement nonequivalence, specifically on those indicators that do not satisfy the invariance constraints, and also allows for further ME/I testing. Despite this fact, there is considerable controversy surrounding the prescription of ‘partial invariance’ testing. The reason for this is twofold: 1) there is no consistency in the statistical criteria used to relax invariance constraints in the literature, and 2) invoking partial invariance constraints have mostly been exploratory and largely capitalize on chance (see Vandenberg & Lance, 2000 for a review). Therefore, in order to ensure that the cross-group comparisons are meaningful and not arbitrary, a majority of the items on a given latent variable should have loadings that are invariant across groups (Reise, Widaman, & Pugh, 1993), and non-invariant items should constitute only a small portion of the model (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000). Within the DIF literature, it is recommended that these non-invariant items be removed from the final test/assessment. Additionally, Vandenberg and Lance (2000) recommended that the items which satisfy ‘partial invariance’ should be selected based on strong theoretical foundations.

2.6.1.4 Invariant unique variance $\Theta_{\delta}^g = \Theta_{\delta}^{g'}$.

The null hypothesis that residual variances are equivalent across groups determines if the scale-items measure the latent constructs with same degree of measurement error (Cheung & Rensvold, 2002; Vandenburg & Lance, 2000). This test is undertaken by constraining like-item uniqueness to be equal between groups (Vandenburg & Lance, 2000). If participants from one (or more) of the group(s) are unfamiliar with the scoring formats of a scale, they are more likely to respond inconsistently to the items (Millsap & Everson, 1993). Furthermore, differences among groups in their vocabulary, grammar, syntax, and common experiences may also produce nonequivalent residual variances (Millsap, 1995).

2.6.2 Tests of Structural Invariance

2.6.2.1 Invariant factor variance $\Phi_{pp}^g = \Phi_{pp}^{g'}$.

Tests the null hypothesis that factor variances are invariant across groups. Factor variances represent the dispersion (or variability) of the latent variable, and therefore this test is frequently treated as complementary to the test of metric invariance (Schmitt, 1982). Differences in factor variances are interpreted as differences in true score calibration across groups. Rejection of the null hypothesis indicates that the group with the smaller factor variance tends to use a narrower range of the construct continuum (Vandenburg & Lance, 2000).

2.6.2.2 Invariant factor covariance $\Phi_{pp'}^g = \Phi_{pp'}^{g'}$.

Tests the null hypothesis that factor covariance are invariant across groups. This test is frequently treated as complementary to the test of Configural invariance. Differences in factor covariances are interpreted as differences in conceptual association of the true scores (Schmitt, 1982; Vandenburg & Lance, 2000). The tests of invariant factor variance and covariance matrices are often combined as an *omnibus test of the equality of the latent variance/covariance matrices* across groups, i.e., $\Phi^g = \Phi^{g'}$ (Byrne, Shavelson, & Muthen, 1989; Vandenburg & Lance, 2000). However, more often than not, the test of invariant factor covariances is not undertaken as a separate test, since most authors (in the review by Vandenburg & Lance, 2000) are of the opinion that there is not much to gain by conducting a test of equality of factor covariance matrices, once the test of configural invariance has been undertaken.

2.6.2.3 Invariant factor means $\kappa^g = \kappa^{g'}$.

Tests the null hypothesis of invariant factor means across groups. This test is analogous to traditional tests of group mean comparisons such as the ANOVA, and begins with an omnibus test of overall group means before moving on to more specific tests (similar to post-hoc tests) to isolate the differences between groups (Cheung & Rensvold, 2002; Schmitt, 1982; Vandenburg & Lance, 2000). The test of invariant factor means, however, is recommended in place of the traditional tests of mean comparisons, since this test corrects for the attenuation of unreliability due to measurement error (Schmitt, 1982),

and also controls for partial measurement nonequivalence by implementing partial invariance constraints (Byrne, Shavelson, & Muthen, 1989).

Although seven different tests for invariance are available within the MG-CFA framework, they are not all used with equal vigor in practice (Vandenberg & Lance, 2000). Overall, there is a general agreement that an omnibus test of the covariance matrices should be undertaken first (Bagozzi & Edwards, 1998; Byrne, Shavelson, & Muthen, 1989; Horn & McArdle, 1992), and many researchers are of the opinion that if the covariance matrices are invariant, no further ME/I testing is required (Bagozzi & Edwards, 1998; Horn & McArdle, 1992; Jöreskog, 1971). Furthermore, Vandenberg and Lance (2000) found that metric and partial metric invariance is most commonly conducted within the MG-CFA literature.

However, from a DIF perspective, both metric and threshold invariance (not commonly conducted within CFA) are relevant. In other words, invariance of the discrimination and difficulty parameters are both tested within the DIF framework. It is, however, possible to easily test for invariance of the threshold parameters within most SEM software applications, such as Mplus, EQS, LISREL, and so on. Therefore, in this study, these two tests of invariance would be given more priority than other tests of invariance. But before we can compare the MG-CFA invariance framework with the IRT-based DIF frameworks, an overview of some of the most commonly used DIF procedures is warranted. Therefore, in the next section, some traditional and more recent DIF detection procedures available within the DIF literature are reviewed.

2.7 COMMONLY USED DIF DETECTION METHODS FOR DICHOTOMOUS AND POLYTOMOUS SCALES

One of the primary goals of DIF analysis is to find items with significant DIF, and exclude them from the final test. It is the hope that the final test no longer demonstrates DIF, and will be a fair assessment measure for members of all racial, ethnic, and gender subgroups (Angoff, 1993; Cole, 1993; Raju, van der Linden, & Fler, 1995; Walker & Beretvas, 2001). Zumbo (2007) points out that DIF detection was largely focused on dichotomous items until the last decade. Moreover, Raju et al. (1995) point out that a psychometric measure of differential functioning for an entire test has been unavailable for a long time. As Camili and Congdon (1999) point out, the “ultimate rationale” behind DIF analyses rests on improving the validity of the score inferences. Rubin (1988) suggests that it would be desirable to have a measure of DIF across items.

Several authors (Camili & Congdon, 1999; Raju, van der Linden, & Fler, 1995; Rubin, 1988) have noted the need for an appropriately defined measure of differential test functioning (DTF), such that the effect of removing or adding items with significant DIF toward the validity of the score inferences from the overall test may be assessed. Raju et al. (1995) maintain that it would be desirable to have an additive property such that individual DIF values sum to the total test DTF for a given set of items. This would mean that if some items had positive DIF (for the *focal group*), and others had negative DIF (for the *focal group*), then the DIF for these two items should cancel out, and the two items together should contribute zero to the overall DTF score for the examinee. They claim that this feature would enable the practitioners to not only assess which items to

delete, but also estimate the net effect of such an action of the overall DTF for the examinee (Oshima, Raju, & Flowers, 1997; Raju, van der Linden, & Fler, 1995).

At the outset, the DIF/DTF research had focused mainly on dichotomously scored items (Angoff, 1993; Camili & Congdon, 1999; Oshima, Raju, & Flowers, 1997; Raju, van der Linden, & Fler, 1995; Zumbo, 2007). However, with the increasing use of performance assessments, the interest in assessing DIF for polytomously scored items has increased (Chang, Mazzeo, & Roussos, 1996; Flowers, Oshima, & Raju, 1999; Gonzalez-Roma, Hernandez, & Gomez-Benito, 2006; Kim & Cohen, 1998; Meade & Lautenschlager, 2004). Furthermore, there is a recent surge of interest assessing the impact of multidimensionality on DIF analysis (Mazor, Hambleton, & Clauser, 1998; Walker & Beretvas, 2001), and some initial attempts at assessing DIF for multidimensional dichotomous models (Oshima, Raju, & Flowers, 1997; Wu & Lei, 2009). Some of the DIF detection techniques that have been most frequently used over the past three decades are summarized below.

2.7.1 Traditional DIF detection procedures

2.7.1.1 ANOVA and ANCOVA-based methods.

ANOVA and ANCOVA-based procedures were initially used to evaluate the extent of *item bias* between sub-groups (Angoff, 1993). Initially ANOVA-based techniques were used in detecting item bias between primarily Black and White examinees performance in SAT examinations (Cardall & Coffman, 1964). Race was introduced as an interaction

term, and this interaction term was tested for significant item bias. However, a statistical implication of the basic conception of DIF required that examinees from the various subgroups be matched on ability before testing for differences in item performance. Therefore, a class of conditional DIF methods based on the ANCOVA technique emerged around 1974. After conditioning on total-score (or examinee ability), these methods studied the effect of the grouping variable(s) and the interaction term(s) on item performance (Angoff & Sharon, 1974). Therefore, these methods were termed the Attribute x Treatment Interaction (ATI) methods (Zumbo, 2007).

2.7.1.2 Delta-plot or Transformed Item Difficulty (TID) method.

Angoff (1972) proposed a method for studying cultural differences using a graphical method based on Thurstone's absolute scaling method (Angoff, 1993). This method, called alternatively as the Delta-plot or TID method, uses the item p-values for the two groups under consideration, and converts these p-values into a normal deviate, expressed on a scale with a mean of 13 and a standard deviation of 4. These pairs of normal deviates for each item are then plotted on a bivariate graph with each group represented on one of the axes. When the groups are of the same level of proficiency, the plot should resemble an ellipse, representing a correlation of .98 or higher (Angoff, 1993). When the groups differ only in proficiency, the plot will be displaced either vertically or horizontally toward the group of higher ability, but the correlation values will not vary significantly (Angoff, 1993). However, when the groups are drawn from different populations the points representing the normal deviates will be dispersed in the off-

diagonal direction, with substantially lower correlation values. Significant outliers that fall outside of the range of other values are said to represent item x group interaction, and are considered exceptionally more difficult for one group than the other (Angoff & Ford, 1973; Angoff, 1993).

2.7.2 Nonparametric DIF detection procedures

The definition of DIF in the nonparametric context is based on observed scores rather than latent scores. These methods typically treat the dependent variables as continuous and examine responses within each score interval or score level (Teresi, 2006a), for example, the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988) and the Standardization procedure (Dorans & Kulick, 1986). Some prominent non-parametric DIF detection procedures are as follows:

2.7.2.1 Mantel-Haenszel (MH) procedure.

The Mantel-Haenszel (MH) χ^2 method (Holland & Thayer, 1988) is probably one of the most widely used DIF detection methods in practice (Teresi, 2006a). This procedure has been used for both dichotomous and polytomous data, and primarily with ordinal polytomous data (Teresi, 2006a). The MH class of methods use contingency tables designed by Mantel and Haenszel (1959) in order to detect DIF (Angoff, 1993; Zumbo, 2007). The contingency table involves two dimensions: (a) student's correct response

[scored (1) for *correct* and (0) for *incorrect*], and (b) group membership [(f) for *focal group* and (r) for *reference group*]. The *reference* and *focal groups* are matched on ability (or total score). The total score is discretized into a number of score category bins or *score intervals* (j). The following is an example of a contingency table as presented in Angoff (1993):

Tested group	Performance on score interval j		
	1	0	
Reference (r)	a_j	b_j	$N_{rj} = a_j + b_j$
Focal (f)	c_j	d_j	$N_{fj} = c_j + d_j$
	$N_{1j} = a_j + c_j$	$N_{0j} = b_j + d_j$	$N_j = a_j + b_j + c_j + d_j$

The MH index is calculated at each *score interval* (j), and is given as follows:

$$\alpha_j = \frac{p_{rj}}{q_{rj}} \bigg/ \frac{p_{fj}}{q_{fj}} = \frac{\frac{a_j}{a_j+b_j}}{\frac{b_j}{a_j+b_j}} \bigg/ \frac{\frac{c_j}{c_j+d_j}}{\frac{d_j}{c_j+d_j}} = \frac{a_j}{b_j} \bigg/ \frac{c_j}{d_j} = \frac{a_j d_j}{b_j c_j} \quad (13)$$

where p_{rj} is the proportion of the *reference group* in *score interval* (j) who answered the item correctly, $q_{rj} = 1 - p_{rj}$, and a_j , b_j , c_j , & d_j refer to the frequencies of *reference* and *focal group* examinees who scored a '1' or '0' on the given item, as represented in the above table. Similarly, p_{fj} and q_{fj} are interpreted for the *focal group*. Therefore, α_j is the ratio (p/q) that represents the odds of getting an item correct for students in the *reference group*, compared to the odds for the students in the *focal group*. If there were no difference in the odds, then α_j will be equal to 1. If, however, the *focal group* performs better than the *reference group* for the said score interval, then $\alpha_j < 1$, and if the *reference group* has better odds of getting the item correct, then $\alpha_j > 1$. The MH procedure estimates a common odds ratio for all matched categories, and the summated MH index is given as follows:

$$\hat{\alpha}_{MH} = \frac{\sum_j p_{rj} q_{fj} N_{rj} N_{fj} / N_j}{\sum_j q_{rj} p_{fj} N_{rj} N_{fj} / N_j} = \frac{\sum_j a_j d_j / N_j}{\sum_j b_j c_j / N_j} \quad (14)$$

which is the average factor by which the odds that a member of the *reference group* responds correctly to the item exceeds the odds that a member of the *focal group* responds correctly to the item (Angoff, 1993). Typically, the score intervals are weighted by the sample size in each group, and intervals where the sample sizes are more equal in the two groups receive higher weights. Additionally, $\hat{\alpha}_{MH}$ is transformed to another scale such that the index is centered around a value of zero (which corresponds to zero DIF). This transformed index is referred to as *MH D-DIF*, which by means of conversion is equal to $-2.35 \ln(\hat{\alpha}_{MH})$. Holland and Thayer (1988) claim that the MH procedure provides both a significance test and a measure of the effect size of $\hat{\alpha}_{MH}$ that is better than comparable chi-square methods (Angoff, 1993).

2.7.2.2 Standardization method.

Dorans and Kulick (1986) proposed a method for identifying DIF which is more or less similar to the MH procedure, and makes use of some of the same information used in the MH index. The correlation between these two indices, when expressed on the same scale, has been found to be .99 or higher (Angoff, 1993). The main differences between the two methods are as follows: (i) the standardization method considers the differences in p -values for the *focal* and *reference group* at each score interval, and (ii) it weights these differences in terms of a specially identified *standardization group*, typically the *focal group*. This weight index is represented as follows:

$$D_{STD} = \sum_j K_j (p_{fj} - p_{rj}) / \sum_j K_j \quad (15)$$

where j refers to the score interval of the matching variable and $K_j / \sum_j K_j$ is the weighting factor at score interval j based on the *standardization group* or *focal group* (Angoff, 1993). Within the framework of the standardization method, the weighting factor K_j could alternatively refer to the number of people in the *reference group*, *focal group*, or total group. The choice of values for K_j varies depending upon investigator preferences, but the number of examinees in the *focal group* is typically used in practice (Dorans & Kulick, 1986).

2.7.2.3 The non-parametric Poly-SIBTEST procedure.

The Poly-SIBTEST is not designed to fit a specific underlying model such as the GRM (Chang, Mazzeo, & Roussos, 1996), and therefore does not make any assumptions about model fit (Bolt, 2002). It however, estimates the differences between groups on expected score, conditional on ability (θ). The expected scores (ES) for the *reference* and *focal group*, $ES_R(\theta)$ and $ES_F(\theta)$ are estimated for the studied item conditioned on total scores for subsets of items (hypothesized to have no DIF). The total scores on these subsets of items (t) are used as proxies for the underlying θ estimate. The $ES_R(t)$ and $ES_F(t)$ for the Poly-SIBTEST are given as:

$$ES_R(t) = \sum_{l=1}^m k P_{Rl}(t) \quad \text{and} \quad ES_F(t) = \sum_{l=1}^m k P_{Fl}(t) \quad (16)$$

where $P_{Rl}(t)$ and $P_{Fl}(t)$ represent the empirical proportion of examinees in each group that obtain score l , and have the valid subscore t . The Poly-SIBTEST incorporates a regression correction procedure which corrects for the measurement error in estimating

the valid subtest, and thereby produces adjusted estimates of the expected scores, $ES_R^*(t)$ and $ES_F^*(t)$. These adjusted estimates are said to reflect examinees of equal ability more accurately across groups. The DIF index for the Poly-SIBTEST framework (denoted as $\hat{\beta}_{UNI}$) uses a weighted average difference of these adjusted expected scores, and is expressed as:

$$\hat{\beta}_{UNI} = \sum_{i=0}^T \left([ES_R^*(t) - ES_F^*(t)] \frac{N_R(t) + N_F(t)}{N} \right) \quad (17)$$

where T is the maximum score on the valid subtest; N is the total number of examinees; and $N_R(t)$ and $N_F(t)$ are the number of examinees obtaining the valid subscore t from each group. $\hat{\beta}_{UNI}$ is said to be approximately normal for large samples, when the null hypothesis of no DIF is assumed. Furthermore, $\hat{\sigma}_{\hat{\beta}_{UNI}}$ is given as:

$$\hat{\sigma}_{\hat{\beta}_{UNI}} = \left[\sum_{i=0}^T \left(\frac{N_R(t) + N_F(t)}{N} \right)^2 \left(\frac{\hat{\sigma}_{Rt}^2}{N_{Rt}} + \frac{\hat{\sigma}_{Ft}^2}{N_{Ft}} \right) \right] \quad (18)$$

The significance test statistic for the Poly-SIBTEST procedure, $SIB = \frac{\hat{\beta}_{UNI}}{\hat{\sigma}_{UNI}}$, is evaluated against a standard normal distribution, and the null hypothesis (of no DIF) is rejected when $|SIB| > z_{1-\frac{\alpha}{2}}$ (Chang, Mazzeo, & Roussos, 1996).

2.7.3 IRT-based DIF detection procedures for dichotomous and polytomous items

In order to compare the performance of examinees from various subgroups, the IRT methods compare the item characteristic curves (ICCs) for a given item computed for members of the *reference* and *focal group(s)*. Within the IRT context, items are said to demonstrate DIF, if the ICCs for the two groups are significantly different. Differences in

ICCs can occur due to *uniform* or *non-uniform* types of DIF referred to earlier. *Uniform DIF* is said to occur when the curves are different in terms of their location or thresholds. In this case, one would assume differences in the difficulty parameter across the two groups. *Non-uniform DIF* is said to occur when the curves are different in their slopes and thresholds, therefore causing differences in difficulty and item discrimination across groups (Angoff, 1993; Teresi, 2006a; Zumbo, 2007). Therefore, the main aim of the IRT-based DIF procedures is to determine the area between the curves for the two groups (Raju, 1990; Zumbo, 2007).

The IRT-based measures do not match the groups on ability or total score, since the IRT parameters are unconditional and the ability function is assumed to be “integrated out”. In other words, the area between the two ICCs is calculated across a continuous latent ability distribution. However, the scale for the latent variable is arbitrary, and the theta (ability) scale must be set during item calibration. This is typically done by setting the mean and standard deviation of the theta distribution to zero and one, respectively. Several authors (Teresi, 2006a; Zumbo, 2007) have pointed out that the family of area measures proposed by Raju and his colleagues (the DFIT measures) are the most commonly used IRT-based DIF procedures. However, the Likelihood Ratio procedure (Thissen, Steinberg, & Gerard, 1986) has also been used fairly frequently (Kim & Cohen, 1998; Stark, Chernyshenko, & Drasgow, 2006).

2.7.3.1 The LR procedure.

The Likelihood Ratio (LR) procedure tests for the differences in estimated IRT item parameters between groups. It was first introduced by Thissen, Steinberg, and Gerard (1986) and uses the Bock and Aitkin (1981) marginal maximum likelihood estimation algorithm. This procedure was extended to the polytomous graded response model by Kim and Cohen (1998). The GRM-LR is the most common form in which the LR test is used (Reise, Widaman, & Pugh, 1993), though the LR test can be used with any IRT model. The LR procedure tests for the differences in IRT item parameters between subgroups by comparing model fit statistics in a series of hierarchical models (Teresi, 2006a). A *compact* model, in which the item parameters for all test items (including items under investigation) are constrained to be equal across groups, is used as the comparative anchor. An *augmented* model is compared to this anchor, and the item parameters (for the items under investigation) are allowed to freely vary in this model (Kim & Cohen, 1998; Thissen, Steinberg, & Gerard, 1986). The LR statistic G^2 , is distributed as a χ^2 distribution, and the null hypothesis is rejected when the calculated G^2 statistic exceeds a critical χ^2 at α . The G^2 statistic is given as:

$$G^2 = [-2\log(compact)] - [-2\log(augmented)] \quad (19)$$

The LR procedure was found to produce consistent Type-I error rates (with nominal α) when the model was not misspecified (Kim & Cohen, 1998; Thissen, Steinberg, & Gerard, 1986).

2.8 DIFFERENTIAL FUNCTIONING OF ITEMS AND TESTS (DFIT) PROCEDURE

Raju, van der Linden, and Fler (1995) first proposed an IRT-based parametric procedure to assess DIF known as the ‘differential functioning of items and tests’ (DFIT) procedure. A need for a measure of overall Differential Test Functioning (DTF) had constantly been pointed out by researchers (Camili & Congdon, 1999; Raju, van der Linden, & Fler, 1995; Rubin, 1988). However, the DFIT was the first method to include such a measure of DTF, which estimates the net effect of items with positive DIF (for the *focal group*) and items with negative DIF (for the *focal group*) on the overall differential functioning of the test (Raju, van der Linden, & Fler, 1995).

Flowers, Oshima, and Raju (1999) extended the initial DFIT framework (Raju et al., 1995) to apply to items that have been polytomously scored in general, but more specifically to apply to the Graded Response Model (GRM). Oshima, Raju, and Flowers (1997) extended the DFIT method for multidimensional dichotomous models. However, these authors have not extended their procedure for multidimensional polytomous models. In this section, the DFIT procedure for unidimensional dichotomous and polytomous items, and the extension of this method for multidimensional dichotomous items proposed by Oshima, Raju, and Flowers (1997) is presented. Subsequently, the logical extension of the DFIT procedure for the multidimensional polytomous models in general, and the MGRM in particular, will be presented.

2.8.1 The DFIT method for dichotomous items

The DFIT procedure starts with an estimation of the differential functioning at the test level (or DTF), and estimates the differential functioning at the item level (DIF) from the covariance between the differential functioning at the item level and at the test level. In order to first present the DFIT procedure for a unidimensional dichotomous model, the two parameter IRT model is briefly presented here. The two parameter logistic (2PL) IRT model (Lord, 1980) for dichotomous item responses generally takes the form:

$$P_i(\theta_s) = \frac{\exp(D\alpha_i(\theta_s - \beta_i))}{1 + \exp(D\alpha_i(\theta_s - \beta_i))} \quad (20)$$

where $P_i(\theta_s)$ represents that probability an examinee s with ability θ might get an item correct; α_i represents the item discrimination for the given item i ; and β_i represents the difficulty of the given item i for examinee s .

The DFIT procedures requires separate item parameter estimation for the *reference group* (R) and the *focal group* (F), resulting in two sets of item parameters for a given test. Then, $P_{iR}(\theta)$ represents the probability for an examinee from the *reference group* with ability θ to get the item correct, and $P_{iF}(\theta)$ represents the probability for an examinee from the *reference group* with ability θ to get the item correct (Raju, van der Linden, & Fler, 1995). The examinee's expected proportion correct (EPC) is expressed summatively within the IRT framework as follows:

$$T_s = \sum_{i=1}^p P_i(\theta_s) \quad (21)$$

where p is the number of items in a test. As pointed out above, within DFIT, two sets of item parameters are estimated for each examinee (one if s/he were a member of the *focal*

group, and one assuming that s/he were a member of the *reference group*). Therefore, each examinee will have two EPCs, T_{sF} and T_{sR} . If $T_{sF} = T_{sR}$, then the examinee's EPC is considered to be independent of group membership. The greater the difference between T_{sR} and T_{sF} , greater the differential functioning of the test. A measure of DTF at the examinee level may then be defined as follows:

$$D^2 = (T_{sF} - T_{sR})^2 \quad (22)$$

In order to generalize this conception of DTF across an entire group (e.g. *focal group*), DTF can also be expressed as the expectation (E) taken across the *focal group* of examinees, and given as:

$$DTF = E_F(D^2) = E_F (T_{sF} - T_{sR})^2 \quad (23)$$

Letting, $D = T_{sF} - T_{sR}$ from Eq. (22), Raju et al., (1995; 1997; 1999) have shown that DTF can also be expressed as integrated over the density function of θ [$f_F(\theta)$] across all examinees in the *focal group* as:

$$DTF = \int_{\theta} D^2 f_F(\theta) d\theta \quad (24)$$

DTF can also be alternatively expressed as:

$$DTF = \sigma_D^2 + (\mu_{TF} - \mu_{TR})^2 = \sigma_D^2 + \mu_D^2 \quad (25)$$

where μ_{TF} is the mean expected score for the *focal group* examinees; μ_{TR} is the mean expected score for the *reference group* examinees (or in other words, the mean expected score for the same group of examinees, but scored as if they were members of the *reference group*); and σ_D^2 is the variance of D . We can derive the DIF for a given item from Eq. (22) as:

$$DIF(d_{is}) = P_{iF}(\theta_s) - P_{iR}(\theta_s) \quad (26)$$

then, DTF may be expressed as:

$$DTF = E \left[\left(\sum_{i=1}^p d_i \right)^2 \right] \quad (27)$$

where, $(d_{is}) = P_{iF}(\theta_s) - P_{iR}(\theta_s)$ and $\sum_{i=1}^p d_i = D$.

Since $DTF = \sigma_D^2 + \mu_D^2$ (Eq. 25), (Eq. 27) can be proved in the following manner:

$$\begin{aligned} DTF &= \sigma_D^2 + \mu_D^2 \\ &= E[(D - \mu_D)^2] + \mu_D^2 \\ &= E(D^2 - 2\mu_D D + \mu_D^2) + \mu_D^2 \\ &= E(D^2) - 2\mu_D E(D) + \mu_D^2 + \mu_D^2 \\ &= E(D^2) - \mu_D^2 + \mu_D^2 \\ &= E(D^2) \\ &= E \left[\left(\sum_{i=1}^p d_i \right)^2 \right] \end{aligned}$$

Further, Raju and colleagues (1995, 1997, 1999) have shown that Eq. (27) can be expressed as a covariance of the difference in expected item scores to the difference in expected test scores, and can be rewritten as:

$$DTF = \sum_{i=1}^p [Cov(d_i, D) + \mu_{d_i} \mu_D] \quad (28)$$

where $Cov(d_i, D)$ is the covariance of the difference in expected item score d_i and the difference in expected test scores D ; and μ_{d_i} and μ_D are the means of the d_i s and the D s respectively. It can be shown that Eq. (28) is directly related to Eq. (27) in the following manner:

Since $D = \sum_{i=1}^p d_i$, then $\sigma_D^2 = \sigma_{d_{i1}}^2 + \dots + \sigma_{d_{iN}}^2 = N\sigma_{d_i}^2$; $E(D) = E(d_{i1}) + \dots + E(d_{iN}) = N\mu_{d_i}$; and $\mu_{d_i} = E(d_i) = \frac{1}{N} \sum_{i=1}^p d_i$. If $DTF = \sum_{i=1}^p [Cov(d_i, D) + \mu_{d_i} \mu_D]$ (Eq. 28), then, we can show that:

$$\begin{aligned}
DTF &= \sum_{i=1}^p [Cov(d_i, D) + \mu_{d_i}\mu_D] \\
&= \sum_{i=1}^p [E(\langle d_i - E(d_i) \rangle \langle D - E(D) \rangle) + \mu_{d_i}\mu_D] \\
&= \sum_{i=1}^p [E((d_i D) - DE(d_i) - d_i E(D) + E(d_i)E(D)) \\
&\quad + \mu_{d_i}\mu_D] \\
&= \sum_{i=1}^p [E((d_i D) - E(D)E(d_i)) + \mu_{d_i}\mu_D] \\
&= \sum_{i=1}^p [E(d_i D) - \mu_{d_i}\mu_D + \mu_{d_i}\mu_D] \\
&= \sum_{i=1}^p [E(d_i D)] \\
&= \sum_{i=1}^p [E(d_i)D] \dots \text{where } D \text{ is a constant} \\
&= \sum_{i=1}^p \left[\frac{1}{N} \sum_{i=1}^p d_i \sum_{i=1}^p d_i \right] \\
&= E \left[\left(\sum_{i=1}^p d_i \right)^2 \right]
\end{aligned}$$

The DFIT framework overall begins with a definition of the DTF and then decomposes into DIF at the item level. Therefore, the definition of DIF would also include information about bias from other items in the test. Raju and colleagues (Flowers, Oshima, & Raju, 1999; Oshima, Raju, & Flowers, 1997; Raju, van der Linden, & Fleer,

1995) referred to this conception [derived from Eq. (28)] as Compensatory DIF or CDIF, and it is expressed as:

$$CDIF_i = Cov(d_i, D) + \mu d_i \mu D \quad (29)$$

where $Cov(d_i, D) = \sigma_{di}^2 + \sum Cov(d_i d_j), i \neq j,$

Combining Eq. (28) and Eq. (29), we can again express DTF as the sum of $CDIF_i$ across examinees and this shows the compensating (or additive) nature of the CDIF. For example, if $P_{iF}(\theta) - P_{iR}(\theta)$ for a given item is -.2, and +.2 for another item, then the bias in favor of the *reference group* in the first item cancels out with the bias in favor of the *focal group* in the second item.

Raju and his colleagues (1995, 1997, 1999) point out that most other measures of DIF assume that all items in the test, other than the item in question, are completely free of DIF. They point out that this is a risky assumption to make, since other items in the test can also contribute to DIF. Furthermore, the additive nature of their CDIF measure is also said to be highly beneficial to the practitioners. It not only enables them to determine which items have CDIF, but it also helps them assess the net effect of deleting an item on the DTF of a test.

Raju and his colleagues (1995; 1997; 1999) also proposed a Non-Compensatory DIF (NCDIF) measure, where DIF is calculated for each item 'i' in question. In this approach, it is assumed that the DIF for all other items would be zero. When DIF is assessed for each individual item 'i' under the assumption that all other items have no DIF, then Eq. (29) can be rewritten as:

$$NCDIF_i = \sigma_{di}^2 + \mu_{di}^2 \quad (30)$$

which does not include information about bias from items other than the studied item 'i'. Raju et al. (1995) showed that their NCDIF measure relates to other IRT-based DIF measures, such as, Lord's (1980) χ^2 measure and Wainer's (1993) DIF measure.

2.8.1.1 Comparison of the CDIF and NCDIF measures.

Raju, van der Linden, and Flier (1995) compared their CDIF and NCDIF measures for dichotomous items. They noted that items having significant NCDIF do not necessarily have significant CDIF, in terms of its effect on the DTF. For example, if one item favors the *reference group*, and another item favors the *focal group*, within the CDIF framework, these items would cancel each other out, and not contribute significantly to the overall DTF. Within the NCDIF framework, however, both these items would still be considered as having significant NCDIF. Therefore, one could end up with a higher number of NCDIF items than CDIF items. In addition to the advantages of cancellation at the test level for the CDIF framework, polytomous items allow for the potential cancellation within an examinee at the item level. It is possible for one category to cancel the effects of another category when computing d_i for a given examinee. For example, if $P_{1iF} > P_{1iR}$, but $P_{2iF} < P_{2iR}$, then they will cancel each other out, leading to a d_i value close to zero (Flowers, Oshima, & Raju, 1999).

Both CDIF and NCDIF can be useful in practice, and the index that is more relevant is dependent on the purpose of the particular study. When total test scores are used for determining the effectiveness of an instructional program, an overall measure of DTF, and therefore the CDIF framework is likely to be more valuable. Furthermore, the CDIF

framework is also useful in selection and placement testing where total score is more interesting. Finally, it is important to consider the net effect on DTF of removing items with compensatory DIF, since it may sometimes not be practical or feasible to remove all items with DIF from the final test. However, when there is concern about the potential offensiveness of a test item to one (or more) of the subgroup(s), the NCDIF approach is likely to be more valuable (Flowers, Oshima, & Raju, 1999; Oshima, Raju, & Flowers, 1997; Raju, van der Linden, & Fler, 1995).

2.8.2 The GRM-DFIT procedure

In general, the GRM has been a very popular model for studying DIF in polytomous items (Reise, Widaman, & Pugh, 1993), since a number of polytomous-DIF detection methods were specifically proposed for the GRM (Kim & Cohen, 1998; Flowers, Oshima, & Raju, 1999). Flowers, Oshima, and Raju (1999) proposed an extension of the original DFIT procedure to apply to polytomous item responses in general, and the GRM in particular.

The GRM (Samejima, 1969; 1972) has been presented in an earlier section in this paper. The general form of the GRM should be recalled from Eq. (1) and Eq. (2), and has been presented here again for clarity purposes:

$$P_{ij}^*(\theta) = \frac{\exp(D\alpha_i(\theta - \beta_{ij}))}{1 + \exp(D\alpha_i(\theta - \beta_{ij}))}$$

where $D = 1.7$; θ represents the examinees' ability level; α_i represents the item slope parameter and β_{ij} represents $k_i - 1$ between category "threshold" parameters (β_{ij}), $j = 1, 2, \dots, k_i - 1$. (Samejima, 1969; 1972).

$$P_{ij}(\theta) = P_{ij}^*(\theta) - P_{i(j+1)}^*(\theta)$$

where $P_{ij}(\theta)$ is the probability of responding in a particular category, conditional on θ . Further, by definition, the probability of responding in or above the lowest category is $P_{i0}^*(\theta) = 1.0$, and the probability of responding above the highest category $P_{ik}^*(\theta) = 0.0$ (Samejima, 1969; 1972).

The calculation of the *expected item-score* (ES_i) demarcates the GRM-DFIT procedure from the dichotomous DFIT procedure (Flowers, Oshima, & Raju, 1999). For the GRM, once the probability of responding in each category is estimated, then the ES_i can be calculated as:

$$ES_i = \sum_{j=1}^k P_{ij}(\theta) x_{ij} \quad (31)$$

where x_{ij} is the score for category j (k is the total number of categories); and P_{ij} is the probability of responding in category j . This is referred to as the *expected item response function*. The *expected item scores* (ES_i) can be summed for an entire test to obtain the *expected test response function* (or T_s) for each examinee, and this is given as:

$$T_s = \sum_{i=1}^p ES_i \quad (32)$$

where p is the number of items in a test. Chang and Mazzeo (1994) demonstrated that if two items have the same ES_i , then they should also have the same number of response categories. However, if the ES_i for a given item is not equal across groups for examinees with a given ability (θ), then the item might be functioning differently for the two groups (Chang & Mazzeo, 1994).

Typically, item parameters are estimated first for the *focal group*, and the estimated item parameters for the *reference group* are then linked on the *focal group* metric using a linear transformation (Flowers, Oshima, & Raju, 1999; Oshima, Raju, & Flowers, 1997; Raju, van der Linden, & Fler, 1995). The *focal group* θ distribution is used to calculate the two ES_i s. Therefore, for any examinee (with a given θ), one expected item score (ES_{iF}) is calculated using the *focal group* item parameters, and another expected item score (ES_{iR}) is calculated using the linked *reference group* item parameters. Therefore, similar to the dichotomous case, if the item is functioning differently, then $ES_{iF} \neq ES_{iR}$. The same reasoning can be applied to the total score on the test, and if $T_{sF} \neq T_{sR}$, then the test is said to be functioning differently for the two groups (Flowers, Oshima, & Raju, 1999; Oshima, Raju, & Flowers, 1997; Raju, van der Linden, & Fler, 1995). Greater the difference between the two expected scores, greater the DTF. Once the two ES_i s for the *reference* and *focal group* are calculated, the DTF and DIF measures for the polytomous models are estimated and interpreted in exactly the same manner as with the dichotomous DFIT measures presented above.

2.8.3 The Multidimensional DFIT procedure for dichotomous items

Oshima, Raju, and Flowers (1997) extended the original DFIT framework for the multidimensional dichotomous items. The multidimensional extension (McKinly & Reckase, 1983; Reckase, 1985; Reckase & McKinley, 1991) of the two-parameter logistic (M2PL) model is given as:

$$P_i(\Theta) = \frac{\exp[D \sum_{h=1}^H \alpha_{ih}(\theta_h - \beta_i)]}{1 + \exp[D \sum_{h=1}^H \alpha_{ih}(\theta_h - \beta_i)]} \quad (33)$$

The DFIT procedure estimates two probabilities for each examinee, one assuming that s/he is a member of the *reference group* (R), and the second assuming that s/he is a member of the *focal group* (F). Therefore, $P_{iR}(\Theta)$ would represent the probability of success on item i for an examinee from the *reference group*; and $P_{iF}(\Theta)$ would represent the probability of success on item i for an examinee from the *focal group*. Once the probability of correct response for an examinee is obtained using the M2PL model, then the examinee total scores, DTF and DIF are obtained in a manner similar to the unidimensional DFIT method.

The main difference in the multidimensional framework, as compared to the unidimensional framework, is the need for multidimensional linking prior to the DIF analysis. In order to compare the two sets of parameters obtained for the *reference* and *focal groups*, the item parameters must be transformed to a common scale. For the multidimensional IRT models, the linking coefficients are obtained by making the following transformations:

$$a^* = (A^{-1})'a \quad (34)$$

$$b^* = b - a'A^{-1}\beta \quad (35)$$

$$\theta^* = A\theta + \beta \quad (36)$$

where A is an $m \times m$ multiplicative linking matrix and β is an $m \times 1$ additive linking vector for the m -dimensional IRT models. The multiplicative linking matrix adjusts variance and covariance differences of ability dimensions for the two groups, and the additive linking vector adjusts the location differences (Oshima, Raju, & Flowers, 1997). Oshima, Davey, and Lee (2000) have shown that the above transformations do not alter

the probability of correct response for an item, and have established model indeterminacy, and a detailed description of this linking procedure can be found in their paper.

This multidimensional linking procedure is an extension of the test characteristic function (TCF) method proposed by Stocking and Lord (1983). The goal is to make ‘a*’ and ‘b*’ in the second group as similar to ‘a’ and ‘b’ in the first group as possible. This done by choosing the correct A and β matrices that would minimize the differences between TCFs. In order to find the minimization function (F_1), equally spaced Θ points are used for matching:

$$F_1 = \frac{1}{L} \sum_{s=1}^L (T_F - T_R)^2 \quad (37)$$

where L represents the equally spaced Θ points in the m -dimensional space. Alternatively, a different minimization function could also be used for the TCF method. The Θ points of the entire *focal group* can be used as matching points, rather than the equally space Θ points called for by Eq. (37). This provides the following minimization function:

$$F_2 = \frac{1}{N_F} \sum_{s=1}^{N_F} (T_F - T_R)^2 \quad (38)$$

It should be recalled from Eq. (22) and (23) that $DTF = (T_{sF} - T_{sR})^2$, and this is exactly what is being minimized in the minimization function. Therefore, the differences between T_{sF} and T_{sR} that are left over after linking, defines DTF and DIF, the differential performance that could be attributed to group membership.

2.8.4 Multidimensional DFIT for polytomous items as an extension of GRM-DFIT

It should be recalled from Eq. (3) above that for a set of H latent traits, the general form of the MGRM is given by:

$$P_{ij}^*(\Theta) = \frac{\exp[D \sum_{h=1}^H \alpha_{ih}(\theta_h - \beta_{ij})]}{1 + \exp[D \sum_{h=1}^H \alpha_{ih}(\theta_h - \beta_{ij})]}$$

where θ_h is the latent trait on dimension h ($h = 1, \dots, H$ dimensions); and the ability level for person 'n' on the H latent traits are represented by a vector of values $\Theta_n = (\theta_1, \dots, \theta_H)^T$; α_{ih} is the discrimination parameter for item i on dimension h ; β_{ij} is the threshold parameter for responding in category j for item i ; D corresponds to a scaling factor of 1.7; and the summation is over all the H dimensions. Within the multidimensional framework, $P_{ij}^*(\Theta)$ is the probability of a randomly selected examinee with latent traits Θ_n responding in category j or higher, for any given item i (De Ayala, 1994).

Additionally, it can be recalled from Eq. (2) that the probability of responding in each category P_{ij} can be easily obtained as follows:

$$P_{ij}(\Theta) = P_{ij}^*(\Theta) - P_{i(j+1)}^*(\Theta)$$

where $P_{ij}(\Theta)$ is the probability of responding in a particular category, conditional on Θ . Further, by definition, the probability of responding in or above the lowest category is $P_{i0}^*(\Theta) = 1.0$, and the probability of responding above the highest category $P_{ik}^*(\Theta) = 0.0$ (Samejima, 1969; 1972).

The DFIT procedure estimates two probabilities for each examinee, one assuming that the examinee is a member of the *reference group* (R), and the second assuming that

the same examinee is a member of the *focal group* (F). Therefore, $P^*_{ijR}(\Theta)$ would represent the probability of success on item i for an examinee from the *reference group*; and $P^*_{ijF}(\Theta)$ would represent the probability of success on item i for an examinee from the *focal group*. Once the probability of correct response for an examinee is obtained using the MGRM model, then, as in the GRM-DFIT procedure, the *expected item-scores* (ES_i s) are calculated (Flowers, Oshima, & Raju, 1999). For the GRM, once the probability of responding in each category is estimated, then the ES_i can be easily calculated as shown in Eq. (31):

$$ES_i = \sum_{j=1}^k P_{ij}(\Theta) X_{ij}$$

where X_{ij} is the score for category j (k is the total number of categories); and P_{ij} is the probability of responding in category j . This is referred to as the *expected item response function*. The *expected item scores* (ES_i) can be summed for an entire test to obtain the *expected test response function* (or T_s) for each examinee, and from Eq. (32) it can be inferred that the T_s for each examinee can be calculated as:

$$T_s = \sum_{i=1}^p ES_i$$

where p is the number of items in a test. Chang and Mazzeo (1994) demonstrated that if two items have the same ES_i , then they should also have the same number of response categories. However, if the ES_i for a given item is not equal across groups for examinees with a given ability (Θ), then the item might be functioning differently for the two groups (Chang & Mazzeo, 1994).

In all the DFIT methods, the item parameters are typically estimated first for the *focal group*, and the estimated item parameters for the *reference group* are then linked on the *focal group* metric using a linear transformation (Flowers, Oshima, & Raju, 1999;

Oshima, Raju, & Flowers, 1997; Raju, van der Linden, & Fleer, 1995). The *focal group* Θ distribution is used to calculate the two ES_i s. Therefore, for any examinee (with a given Θ), one expected item score (ES_{iF}) is calculated using the *focal group* item parameters, and another expected item score (ES_{iR}) is calculated using the linked *reference group* item parameters. Therefore, similar to the dichotomous case, if the item is functioning differently, then $ES_{iF} \neq ES_{iR}$. The same reasoning can be applied to the total score on the test, and if $T_{sF} \neq T_{sR}$, then the test is said to be functioning differently for the two groups (Flowers, Oshima, & Raju, 1999; Oshima, Raju, & Flowers, 1997; Raju, van der Linden, & Fleer, 1995). Greater the difference between the two expected scores, greater the DTF. Once the two ES_i s for the *reference* and *focal group* are calculated, the DTF and DIF measures for the polytomous models are estimated and interpreted in exactly the same manner as with the dichotomous DIFT measures presented above.

However, in order to compare the two sets of parameters obtained for the *reference* and *focal groups*, the item parameters must be transformed to a common scale, and multidimensional linking for the MGRM parameters has to be performed. For the multidimensional IRT models, the linking coefficients are obtained by making the following transformations:

$$a^* = (A^{-1})'a \quad (39)$$

$$\bar{B}^* = \bar{B}_R - (a' A^{-1} \beta) \quad (40)$$

$$\theta^* = A\theta + \beta \quad (41)$$

where A is an $m \times m$ multiplicative linking matrix, β is an $m \times 1$ additive linking vector for the m -dimensional IRT models, and where \bar{B}_R is the mean of the β_{ijR} s. The multiplicative linking matrix adjusts variance and covariance differences of ability

dimensions for the two groups, and the additive linking vector adjusts the location differences (Yao, 2004b; Yao & Boughton, 2007; 2009). Oshima, Davey, and Lee (2000) have shown that the above transformations do not alter the probability of correct response for an item, and have established model indeterminacy, and a detailed description of this linking procedure can be found in their paper.

This multidimensional linking procedure is an extension of the test characteristic curve (TCC) method proposed by Stocking and Lord (1983). The goal is to make ‘a*’ and ‘ \bar{B}^* ’ in the second group as similar to ‘a’ and ‘ \bar{B} ’ in the first group as possible. This done by choosing the correct A and β matrices that would minimize the differences between TCCs. The minimization function (F) is obtained for equally spaced Θ points by using an extension of Stocking and Lord’s (1983) Test Characteristic Curve (TCC) method for the GRM provided by Baker (1992). In Baker’s proposed technique, the two equating coefficients are obtained by minimizing the quadratic loss function (F):

$$F = \frac{1}{N} \sum_{s=1}^N (T_{sR} - T_{sF})^2 \quad (42)$$

where N is the arbitrary number of equally spaced points along the Θ metric, T_{sR} and T_{sF} are the true scores for the *reference* and *focal* groups respectively, and are defined as:

$$T_{sR} = \sum_{i=1}^p \sum_{j=1}^k u_{ij} P_{ijR}(\Theta_s) \quad (43)$$

$$T_{sF} = \sum_{i=1}^p \sum_{j=1}^k u_{ij} P_{ijF}(\Theta_s) \quad (44)$$

where u_{ij} is the weight allocated to response category ‘j’ for item ‘i’, and is typically the integer index for that category. The main task in the TCC method is to find the values for the A and β transformation matrices that minimize the quadratic loss function given in Eq. (39). Multidimensional linking for the GRM has been incorporated in the LinkMIRT software (Yao, 2004b; Yao & Boughton, 2007; 2009). Once the parameters are linked

using the Stocking and Lord Method, then the ESs can be calculated as presented above, and the rest of the DFIT framework follows the same logic as presented in the unidimensional case.

2.8.5 DFIT Significance Tests

If the D between expected scores is assumed to be normally distributed, with a mean of μ_D and a standard deviation of σ_D . A Z score for examinee s can be calculated as:

$$Z_s = \frac{D_s - \mu_D}{\sigma_D} \quad (45)$$

where Z_s^2 has a χ^2 distribution with one degree of freedom (df). The sum of Z_s^2 across all examinees has a χ^2 distribution with N_{df} , and is given as:

$$\chi_N^2 = \sum_{s=1}^N Z_s^2 = \frac{\sum_{s=1}^N (D_s - \mu_D)^2}{\sigma_D^2} \quad (46)$$

The object is to minimize the expectation of the DTF, which would mean that μ_D^2 should resolve to zero. Then Eq. (46) could be re-expressed as:

$$\chi_N^2 = \frac{\sum_{s=1}^N D^2}{\sigma_D^2} = \frac{N(DTF)}{\sigma_D^2} \quad (47)$$

A significant χ^2 value would indicate that one or more items have DIF. Raju and colleagues (1995; 1997; 1999) recommend that practitioners should begin by removing items with significant CDIF until the χ^2 value is no longer significant. The χ^2 test for NCDIF, however, was shown to be extremely sensitive to large sample sizes, and it has been recommended that a critical value be empirically established for NCDIF (Flowers,

Oshima, & Raju, 1999; Oshima, Raju, & Flowers, 1997; Raju, van der Linden, & Fleer, 1995).

2.9 PERFORMANCE OF THE DFIT METHODS AND COMPARISONS TO OTHER DICHOTOMOUS AND POLYTOMOUS DIF DETECTION PROCEDURES

Raju, van der Linden, and Fleer (1995) first proposed the DFIT measures for unidimensional dichotomous items in the seminal paper, and this framework has since then been widely used in DIF research (Teresi, 2006a; Zumbo, 2007). The DFIT family of measures provides an approach for assessing DIF both at the item level and at the test level (Morales, Flowers, Gutierrez, Kleinman, & Teresi, 2006; Raju et al., 1995; 1997; 1999). The DFIT indices reflect the overall magnitude of DIF in addition to identifying items showing DIF, which provides users some guidance with regard to the impact an item has on the overall scale differential functioning (Morales, Flowers, Gutierrez, Kleinman, & Teresi, 2006). Unlike other IRT-based methods, such as Lord's (1980) χ^2 test and Raju's (1990) signed area measure (which are based on a theoretical range of θ), the DFIT measures are based on the actual distribution of the ability estimates within the group for which DIF is estimated (Morales, Flowers, Gutierrez, Kleinman, & Teresi, 2006). However, the procedure to implement the DFIT is complicated, requiring 3 separate computer programs, for estimating the item parameters (such as BILOG, MULTILOG, or PARSCALE), for equating the estimated item responses (such as EQUATE, IPLink, or LinkMIRT), and the DFIT software (which is not freely accessible

to researchers). Additionally, the significance tests and empirically derived cutoff values for the DFIT tests are often debated as sensitive (Morales, Flowers, Gutierrez, Kleinman, & Teresi, 2006). Therefore, a review of the performance of the DFIT measures in assessing DIF as against other IRT-based methods is warranted.

Raju et al. (1995) compared their CDIF and NCDIF measures for dichotomous items with Lord's (1980) χ^2 test and Raju's (1990) signed area measure. Data were generated to simulate four proportions of test-wide DIF (0%, 5%, 10% and 20%) for a 40 item test. Furthermore, the direction of DIF was manipulated in this study such that, in the *Unidirectional* DIF conditions, items always favored the *reference group*; while in the *Balanced-bidirectional* DIF conditions, items favoring the *reference group* were balanced with items favoring the *focal group*. Equal numbers of *uniform* and *non-uniform* DIF items were introduced across conditions. In addition, sample size was varied at two levels (N=500 and N=1000) across all conditions. The item parameters for the two groups were linked using the Stocking and Lord (1983) TCF method. Item parameter recovery (using RMSDs and correlation), and Type-I error rates were assessed.

Raju et al. (1995) found that the DFIT measures were much more accurate in detecting DIF and DTF with relatively few detection errors (false positives) across all simulated conditions. The most number of false positives and false negatives were found in the N=500 condition with 20% DIF items. The number of false negatives declined as sample size increased. Raju's signed area measure, however, did not perform as well as the DFIT measures (CDIF and NCDIF) and Lord's χ^2 measure with respect to false positives. However, with respect to false negative identification, they found no differences in the performance of the signed area measures and Lord's χ^2 measure, while

the CDIF and NCDIF measures performed significantly better in minimizing the number of false negatives identified (Raju, van der Linden, & Fler, 1995).

Flowers, Oshima, and Raju (1999) proposed a DIF detection procedure for the GRM based on the DFIT (Raju, van der Linden, & Fler, 1995) family of measures. In order to assess the robustness of the GRM-DFIT procedure, Flowers, Oshima, and Raju compared the accuracy of DIF detection using this measure across various study conditions. They simulated data for the *focal* and *reference groups* under an *Equivalent θ* condition, where both groups were sampled from a $N(0, 1)$ θ distribution. They also used a *Non-Equivalent θ* condition, where the *focal group* was sampled from a $N(-1, 1)$ θ distribution, resulting in a series of conditions where the *focal group* of examinees would have lower ability than the *reference group*. Both *Unidirectional* and *Balanced-bidirectional* DIF conditions were used. Additionally, they simulated data under two test length (20 items and 40 items) conditions, with four proportions of test-wide DIF (0%, 5%, 10%, 20%) for each test length. For the 20 item study conditions with 5% DIF, only one item would be embedded with DIF, and therefore, the *Bidirectional* DIF condition was not simulated for these conditions.

In their previous work, Raju and colleagues (Oshima, Raju, & Flowers, 1997; Raju, van der Linden, & Fler, 1995) found that the NCDIF measure was extremely sensitive to large samples, and rejected the null hypothesis very frequently. Therefore, an empirical critical value (of .016) was established for the NCDIF measure, by finding the 99th percentile value for DIF analysis conducted on 2000 DIF-free items. Flowers et al. (1999) calculated two indicators to determine the accuracy of DIF detection using the CDIF and NCDIF measures. A *true positive (TP)* indicator was used to estimate an embedded DIF

item with a DIF index above the cutoff value. A *false positive (FP)* indicator was used to estimate a non-DIF item with a DIF index above the cutoff value. It was expected that higher *TP* and lower *FP* estimates would reflect positively on the robustness of these DIF measures across conditions.

Overall, the polytomous GRM-DFIT procedure was found to be effective in identifying DTF and DIF. In general however, they found that it was easier to detect larger amounts of DIF in items than small amounts of DIF (Flowers, Oshima, & Raju, 1999), and some items with extremely small embedded DIF values ($< .10$) were undetected across all estimated replications. In addition, Flowers et al. (1999) did find that CDIF was not as stable as NCDIF. For the CDIF measure, they found that the proportion of *false positives* were larger under a couple of conditions with 20-items. They attributed the occasional inconsistencies in CDIF estimation to the fact that linking errors associated with each item tended to accumulate across the entire test (since CDIF values were summed across the entire test). Furthermore, since DTF was estimated directly from CDIF, the DTF estimates also tended to be unstable compared to the NCDIF measures. However, since NCDIF was estimated uniquely for each item, it was more stable in estimating DIF across conditions. In general, they found that the NCDIF measure resulted in larger *TP* estimates and almost no *FP* estimates across all study conditions.

On the whole, barring the occasional erratic *FP* detection rate for the CDIF measure, Flowers et al. (1999) found that both their CDIF and NCDIF measures were robust in detecting DIF across most of their study conditions for data simulated under the graded model. However, with real data, any model utilized will only be an approximation of the “true” underlying response patterns (Bolt, 2002). Since both the GRM-LR and the GRM-

DFIT are parametric procedures, developed specifically for the Graded model, they are expected to be highly model-dependent. Therefore, Bolt (2002) points out the practical importance of assessing the robustness of parametric methods to model misfit, specifically the degree to which item parameter invariance is preserved when the underlying model is misspecified. Furthermore, Flowers et al. (1999) do recognize the fact that their findings are limited to the conditions tested in their study, and for when the model is accurately identified. They, therefore, urge researchers to examine the Type I and Type II errors in estimating DIF and DTF using the GRM-DFIT procedure, as against other polytomous DIF detection methods (Flowers, Oshima & Raju, 1999).

Bolt (2002) compared the parametric procedures of GRM-DFIT and GRM-LR to the nonparametric Poly-SIBTEST procedure under conditions when the model was accurately specified (and fit the model), and when there was small amounts of model misfit. Model misfit was simulated by applying the GRM-based DIF procedures to data generated either from a Generalized Partial Credit Model (GPCM) or from a 2p-Sequential Response Model (SRM). He claimed that both these models provide mechanisms for data simulation that were statistically different from the GRM, but difficult to distinguish from the GRM based on goodness-of-fit to the model. Both the Type-I error rate and power of these three DIF detection procedures were compared (Bolt, 2002).

For both the Type-I error and power studies, the same item parameters were used to generate the data for the *reference* and *focal groups*, but the underlying simulation model were varied (GRM, GPCM, or the 2p-SRM) across conditions. Additionally, the *focal group* item parameters were varied in order to reflect the types and amounts of DIF.

Sample size and the latent mean difference between the subgroup ability distribution were also varied ($N = 300$ and $N = 1000$; $\mu_D = 0$ and $\mu_D = 1$). Bolt (2002) found that when the generating model was the GRM, Type-I errors were consistent with the nominal- α . Furthermore, when the mean ability difference between the *reference* and the *focal groups* was zero, $\mu_D = 0$, the GRM-LR procedure still produced Type-I errors close to the nominal- α , even when the data was generated under the GPCM or the 2p-SRM models. However, under this procedure, when $\mu_D = 1$, the Type-I error rate was mostly inflated for almost all items when the generating model was misspecified (Bolt, 2002).

With the GRM-DFIT procedure, a couple of items were more susceptible to Type-I inflation under model misspecification, even when $\mu_D = 0$. This was because the same empirical critical values were applied across all items for the NCDIF measure. But when it came to the $\mu_D = 1$ conditions, again the GRM-DFIT procedure performed equally well, and the Type-I error rate was not highly inflated for most items (except the couple that were already affected in the $\mu_D = 0$ condition) (Bolt, 2002). For the non-parametric Poly-SIBTEST procedure, Bolt (2002) found that the Type-I error results were consistently close to the nominal- α under all study conditions. The tables however, were completely reversed for the results from the power study. The power rates for both the GRM-LR and the GRM-DFIT procedures were almost unaffected by model misspecification, while the non-parametric Poly-SIBTEST produced noticeably lower power rates, than the other two procedures, in almost all study conditions (Bolt, 2002).

Therefore, the parametric tests are definitely able to detect DIF with greater power than the non-parametric tests. Even though non-parametric tests might have some advantages when it comes to Type-I error inflation, these methods tend to lack power

(Chang, Mazzeo, & Roussos, 1996; Bolt, 2002). Furthermore, Bolt (2002) found that the non-parametric method performed significantly better than the GRM-LR method, but its advantages in controlling for Type-I inflation were not necessarily significant over the GRM-DFIT procedure. Overall, the GRM-DFIT was found to be much less affected by model misfit than the GRM-LR method (Bolt, 2002), and it also demonstrated much higher power than the non-parametric Poly-SIBTEST procedure (Bolt, 2002). Therefore, the GRM-DFIT procedure appears to be a powerful and robust test for detecting DIF for polytomous test items (Flowers, Oshima, & Raju, 1999; Bolt, 2002).

The NCDIF measure relies on non-parametric cutoff values in order to evaluate DIF, and despite constant debate over the optimal cutoff values for NCDIF, no empirical simulation was conducted to provide users alternate cutoff values for the NCDIF measure (Bolt, 2002). Therefore, in order to fill this gap in DFIT application, Meade, Lautenschlager, and Johnson (2006) evaluated a number of alternate cutoff values for the NCDIF measure by simulating data for a 12-item polytomous Likert-type (5-point) measure using the GRM-DFIT procedure. Three kinds of uniform DIF conditions were used in this study: only varying the largest 'b' parameter across groups, wherein the most extreme option (5) would be more likely used by the *reference group*; varying the two largest 'b' parameters across groups, wherein the options of (4) and (5) would be more likely used by the *reference group*; and finally varying all 'b' parameters across groups.

In addition, two sample size conditions (500 and 1000), and two magnitudes of DIF (large, 1.0; and small, 0.4) were also simulated in this study (Meade, Lautenschlager, & Johnson, 2006). Finally, bi-directional DIF was simulated such that DIF cancelled out for all items when the overall DTF is estimated (Meade, Lautenschlager, & Johnson, 2006).

Seven alternative cutoff values were examined for the NCDIF measure in this study. They investigated the cutoff value of .096 (recommended by Raju et al., 1995), along with cutoff values of .054 (recommended if one response option were deleted due to low response rate), .032 (recommended by Bolt, 2002), .016 (recommended by Flowers et al., 1999), and three other empirically derived cutoff values of .0115, .009, and .006. They used ROC curves to evaluate Power and Type-I error rates for each NCDIF cutoff value (Meade, Lautenschlager, & Johnson, 2006).

They found that higher cutoff values were associated with both lower power and lower Type-I error across all study conditions. Cutoff of .096 produced the lowest power, while the number of *false positives* was almost close to zero in all conditions. On the other hand, the cutoff value of .006 produced very high Type-I error rates, especially for a sample size of 500. Overall, they found that large amounts of DIF are detected by any NCDIF cutoff value (with values between .054 and .009 all performing equally well), and none of the cutoff values exhibited adequate power while maintaining adequate Type-I error rate for very small amounts of DIF. Therefore, the question of choosing an optimal cutoff was relevant only for moderate amounts of DIF. When a moderate amount of DIF was present, a cutoff value of .0115 and .009 performed most optimally for sample sizes of 500, while cutoff values of .009 and .006 produced the most optimal balance of Power and Type-I error rates, for sample sizes of 1000. Overall, they recommend that researchers derive their own cutoff values empirically based on the number of items and sample size, whenever possible (Meade, Lautenschlager, & Johnson, 2006).

2.10 CFA-BASED APPROACHES FOR DIF DETECTION AND SOME COMPARISON WITH IRT-BASED DIF DETECTION

A number of studies have recently been exploring DIF detection from the MG-CFA perspective (Glockner-Rist & Hoijtink, 2003; Gonzalez-Roma, Hernandez, & Gomez-Benito, 2006; Meade & Lautenschlager, 2004a; 2004b; Raju, Laffitte, & Byrne, 2002; Stark, Chernyshenko, & Drasgow, 2006), and some of these studies have also compared the MG-CFA approach to other IRT-based approaches to DIF detection, such as the DFIT method (Raju, Laffitte, and Byrne, 2002) and the IRT-LR approach (Stark, Chernyshenko, & Drasgow, 2006). The two approaches (IRT and CFA) have a number of underlying similarities (Takane & de Leeuw, 1987) which warrant a comparison of the two approaches in general, and the DIF analyses within the two approaches in particular. A number of authors (Glockner-Rist & Hoijtink, 2003; Raju, Laffitte, and Byrne, 2002; Stark, Chernyshenko, & Drasgow, 2006) have clearly pointed out the similarities and differences between the CFA and IRT perspectives for DIF assessment.

Both the CFA and IRT perspectives examine the relationship between an underlying construct and a set of measured variables. However, the nature of the relationship between the latent construct and the item true score is different in both approaches. While a linear relationship between the latent and observed variables is assumed in the CFA framework, the underlying relationship is assumed to be nonlinear within the IRT framework. The probability of responding to a dichotomous or polytomous item is expressed as a logistic function in IRT, and is therefore a nonlinear function. However, the logarithm (\ln) of an odds ratio [$\ln(P/(1-P))$] is linear even in the IRT framework (Raju, Laffitte, & Byrne, 2002).

Both CFA and IRT approaches examine the degree to which this relationship holds for persons from different populations (or sub-populations), who have the same level of the underlying construct of interest (be it ability, satisfaction, or attitude). Both perspectives do not assume that the distribution of scores on the underlying construct is identical in the populations being compared. The latent mean difference between the two groups is typically referred to as '*impact*'. On the other hand, both perspectives, by definition, imply that persons with identical latent scores will have identical true scores on the item and test level, irrespective of the group they belong to (Raju, Laffitte, & Byrne, 2002). When the measurement equivalence between the two groups cannot be established, both the CFA and the IRT frameworks can be used to identify the extent and the source of the problem. However, the two frameworks differ in their approaches here.

In MG-CFA, metric invariance is cited as a prerequisite for meaningful examination of threshold invariance (Vandenberg & Lance, 2000), whereas in IRT analysis, simultaneous tests of DIF in both the discrimination and difficulty indices are performed. Comparisons of uniqueness invariance and item reliabilities, which are routinely performed within the CFA framework, are not typical to IRT applications (Stark, Chernyshenko, & Drasgow, 2006). Instead, in the IRT applications, item and test information functions obtained from standard errors (conditional on θ) are compared (Stark, Chernyshenko, & Drasgow, 2006). Even though, the standard error functions could be integrated and used to compute reliabilities (Green, Bock, Humphreys, Linn, & Reckase, 1984), statistical tests of equal reliability are not routinely performed in relation to measurement equivalence within the IRT framework. Finally, while individual items are specifically assessed for DIF in the IRT approach, the proposed model is tested for its

goodness of fit to the data within the CFA perspective. However, once the nonequivalence is established, items are examined for remedial purposes in MG-CFA as well (Raju, Laffitte, & Byrne, 2002).

The IRT-LR method is an exception to typical IRT-based approaches, in that this method uses a series of nested hierarchical models for comparison, in a manner similar to the CFA perspective (Kim & Cohen, 1998). However, the pattern of model testing is again different between this IRT-based method, and the MG-CFA method. In the IRT-LR method, all parameters other than the referent (or studied) item are fixed to be equal across groups in the baseline model. In each subsequent model, all items other than the one currently being studied are constrained to be equal. However, within the MG-CFA framework, quite the opposite approach is taken. In the baseline model, all parameters other than those for the referent item are free to vary, and only the studied item is constrained to be equal across groups. If the chi-square test is not significant, then the next item is added to the constraints. Subsequent models thus add items to be constrained to be compared to an initially unconstrained model (Stark, Chernyshenko, & Drasgow, 2006).

Despite some of these aforementioned differences, it becomes obvious that the two approaches (IRT and CFA) are unified in their purpose of estimating the relationship between an underlying trait and an observed variable, and are interested in establishing this relationship across groups. Moreover, the methodologies embodied in both approaches are largely comparable (Takane & de Leeuw, 1987) and the statistical equivalence between the two methods has been established earlier in this paper. A review

of some of these studies comparing IRT and CFA approaches to DIF detection is provided here.

Raju, Laffitte, and Byrne (2002) used a 10-item scale (with 5 response categories) measuring satisfaction in work assignment taken from the Armed Forces Sexual Harassment Survey (Collins, Raju, & Edwards, 2000). They used the GRM-DFIT approach and the MG-CFA approach to examine DIF between a sample of Black (N=1000) and White (N=1000) active duty personnel in the Army, Navy, Marine Corps, and Coast Guard. After establishing the goodness-of-fit of the model for both the Black and White samples, the authors tested for the equivalence of the scale items between groups using the MG-CFA perspective. In testing for equivalence, they first compared a model in which all factor loadings for the items were constrained to be equal across groups to a baseline model where all factor loading were freely estimated. This model comparison resulted in a significant χ^2 difference test, and therefore they sought to identify individual invariant items by constraining a single item in each subsequent model. They found that eight of the ten items (other than item 1 and 2) were equivalent between the groups when investigated from a CFA perspective (Raju, Laffitte, & Byrne, 2002).

In order to replicate these findings they used the GRM-based DFIT (only NCDIF was used in this case) measures (Flowers, Oshima, & Raju, 1999). The item parameters for the *reference group* (White) were equated to the same scale as the parameters underlying the *focal group* (Black). Using the DFIT measures, they found that only one item (item 2) had significant NCDIF across groups, and the DTF across groups was not significant. They found that this item favors the White respondents, for example, a person

with a satisfaction level of $\theta = 2$, might choose a category four (agree) response if they were Black, but choose a category five (strongly agree) response if they were White.

They claim that the CFA and IRT perspectives are comparable since item 2 showed up as nonequivalent in the CFA analysis and as having significant DIF in the DFIT analysis. Additionally, eight other items (items 3 to 10) showed up as invariant in both analyses. However, while item 1 was marked as nonequivalent between the groups in the CFA analysis, neither did item 1 show up as significant in the NCDIF analysis, nor did the IRFs show any marked differences between the two groups. They conjecture that the CFA method might pick up more instances of measurement nonequivalence than the IRT method, since the CFA method is linear. However, it is unclear which of these approaches is more accurate in its analysis and they propose that a comprehensive Monte Carlo study should be undertaken to substantiate this hypothesis (Raju, Laffitte, & Byrne, 2002).

Meade and Lautenschlager (2004a) did exactly this in their simulation study comparing DIF from CFA and IRT perspectives. Their main thesis underlies the assumption that only one intercept parameter (τ_p) is estimated under the CFA framework per item, and therefore, when compared to the IRT method, differences in the 'b' parameter across groups (or *uniform* DIF) will not be picked up as well in the CFA method. However, they claim that differences in the 'a' parameter (or *non-uniform* DIF) will be picked up equally well under both the IRT and CFA methods. It has to be pointed out here that the between-category threshold parameters (τ_{ij}), analogous to the (β_{ij}) parameters in IRT, can also be estimated from a CFA-perspective (see Kannan & Kim, 2009).

Therefore, the purported advantages/disadvantages of estimating only one intercept parameter in CFA would not hold when the between-category thresholds are individually assessed in the CFA-framework, and estimating the ' τ_{ij} ' parameters allows for a more straightforward comparison of IRT and CFA estimation procedures. Nevertheless, Meade and Lautenschlager (2004a) conducted one of the first empirical simulation studies comparing the IRT-LR method and the MG-CFA methodologies, and presenting findings from their study is relevant in terms of discussing future studies conducted in comparing these two estimation methods.

Meade and Lautenschlager (2004a) simulated data for a six-item 5-point scale measuring a single construct. In addition, the following variables were manipulated in this study: sample size (150, 500 and 1000); number of DIF items (2 and 4); and type of DIF ('b' parameter or *uniform* DIF, and 'a' parameter or *non-uniform* DIF). Three kinds of uniform DIF conditions were used in this study: only varying the largest 'b' parameter across groups, wherein the most extreme option (5) would be more likely used by the *reference group*; varying the two largest 'b' parameters across groups, wherein the options of (4) and (5) would be more likely used by the *reference group*; and finally varying the two extreme 'b' parameters across groups, wherein the options of (1) and (5) would be more likely used by the *reference group* (Meade & Lautenschlager, 2004a).

Meade and Lautenschlager (2004a) found that, for the *uniform* DIF conditions, the MG-CFA omnibus test of MI was largely inadequate at detecting DIF, especially when sample sizes were 150. With sample sizes of 500 and 1000, *uniform* DIF was more easily identified in items, but the source of the DIF (i.e., identifying the specific category, (1), (4), or (5)) was not possible within the CFA analysis. However, the IRT-LR method was

more successful in identifying *uniform* DIF more accurately in items, and also identifying the source of the DIF in terms of the exact category where DIF lies. Nevertheless, the IRT-LR method was not able to detect any DIF for the 150 sample size condition (Meade & Lautenschlager, 2004a).

For the *non-uniform* DIF conditions as well, they found that the IRT-LR method performed better than the CFA method in detecting DIF items. However, the performance of both methods in detecting DIF was largely dependent on sample size, and for the N=1000 condition, the IRT-LR method was able to accurately detect all DIF items, while the CFA method was still unable to detect some of the items. Furthermore, for N=500, the omnibus ME/I test under CFA picked up any differences across samples only for 25 of the 100 replications, whereas the IRT-LR method was more accurate in detecting DIF items at this sample size for *non-uniform* DIF (Meade & Lautenschlager, 2004a).

Therefore, Meade and Lautenschlager (2004a) not only found that the IRT-LR method was more accurate for detecting DIF in the ‘b’ parameter, they also found that the IRT-LR method was more accurate in detecting DIF for the ‘a’ parameter as well. Based on these results, they question the analogy between IRT and CFA parameters. These findings are quite contrary to what would generally be expected from the MG-CFA approach, especially given that these authors found the IRT method to be more robust even at smaller sample sizes. It has been found in several subsequent studies that IRT-based estimation methods require larger sample sizes, and that CFA methods are more robust at smaller samples (Stark, Chernyshenko, & Drasgow, 2006; Gonzalez-Roma, Hernandez, & Gomez-Benito, 2006; Wu & Lei, 2006), and therefore a review of these studies are now presented.

Stark, Chernyshenko, and Dransgow (2006) compared DIF detection using both the mean and covariance structures (MACS) CFA-based method, and the IRT-based LR method. They claimed that it is important to be consistent in the methodologies used, in order to be able to compare IRT and CFA-based methodologies to DIF testing, and point out that earlier studies (e.g., Meade & Lautenschlager, 2004a) have not done this. Stark, et al. reiterated the fact that the IRT and CFA methodologies differ in their hierarchical modes of comparison, and while one (IRT) method uses a constrained baseline model, the other (CFA) uses a free baseline model. Therefore, in order to be consistent in their methodology and be able to compare the two approaches, they proposed a common strategy for DIF detection using a free baseline model and a Bonferroni corrected critical-p value under both approaches (Stark, Chernyshenko, & Dransgow, 2006).

Stark et al. (2006) wanted to be able to compare the performances of both free baseline and constrained baseline model approaches, and therefore used both kinds of baseline models under both CFA and IRT analysis. They simulated a 15-item unidimensional scale, and compared the Type-I error and power rate for the two methods of DIF analysis (*MACS vs. IRT-LR*) across the following manipulated variables: amount of DIF (*no DIF; small DIF, $\lambda_{FG} < by .15$, $\tau_{FG} < by .25$; and large DIF, $\lambda_{FG} < by .4$, $\tau_{FG} < by .5$), type of DIF (*uniform vs. non-uniform*), amount of impact (*equal latent mean distribution vs. moderate impact $\mu_{RG} > \mu_{FG}$ by 0.5*), number of response categories (*dichotomous vs. polytomous*), sample size (*500 vs. 1000*), type of baseline model (*free vs. constrained*), and the critical *p*-value for rejecting the null (*.05 vs. a Bonferroni corrected critical-p of .05114*). Additionally, the same critical χ^2 (2) values ($\chi^2 = 5.99$), with 2 degrees of freedom, were used for both the dichotomous and polytomous MACS*

analyses. This is because the number of estimated parameters does not change for MACS analyses. However, the polytomous IRT analysis required the estimation of three additional location parameters (for a 5-option data), and therefore the critical χ^2 values were based on 5 degrees of freedom, ($\chi^2 = 11.07$). Again, it has to be pointed out that only a single intercept value (τ_p) was estimated in the CFA analysis here, and the between-category threshold parameters (τ_{ij}), analogous to the (β_{ij}) parameters in IRT, which can be estimated from a CFA-perspective (see Kannan & Kim, 2009), have not been estimated separately here.

Stark, et al. (2006) found that, overall in their no-DIF conditions, MACS and IRT-LR performed comparably across all study conditions. The Type-I error rate increased slightly as sample size increased, but was not affected by the amount of impact. They found that, for the no-DIF conditions, Bonferroni corrections almost eliminated Type-I errors (producing Type-I error rates of .01 at the maximum). However, when DIF was introduced, the results varied depending upon the number of categories (dichotomous or polytomous), amount of DIF (low vs. high DIF), type of DIF (uniform vs. non-uniform) and type of baseline model used. For the dichotomous (small DIF) conditions, they found that power varied with sample size for the non-uniform DIF conditions, but not for the uniform DIF conditions (power was consistently high in these conditions). However, for the uniform DIF conditions (dichotomous case), the Type-I error rate was inflated across all conditions, but was much lower for the non-uniform DIF conditions. On the other hand, for the large DIF conditions, they found that power was high (1.00) across the board, but the Type-I error rate was also markedly increased across all large DIF conditions (Stark, Chernyshenko, & Drasgow, 2006).

Overall, the most important finding for the dichotomous models noted by Stark et al. (2006) was that the free baseline models produced significantly lower Type-I error rates than the constrained baseline models. However, using Bonferroni corrections significantly improved the inflation of Type-I errors for the dichotomous conditions, even in the case of the constrained baseline model conditions, but almost eliminated them for the free baseline conditions. However, the Bonferroni corrections tended to adversely impact the power of DIF detection, for the small DIF, but not the large DIF conditions. Finally, for the dichotomous models overall, they did not find a significant difference between MACS and IRT-LR in their power and Type-I error detection (Stark, Chernyshenko, & Drasgow, 2006).

However, Stark et al. (2006) point out that for the polytomous response conditions, even though the Type-I error rate did not differ for the two estimation methods, the MACS method had significantly higher power when compared to the IRT-LR method in detecting DIF items, especially when the sample size was low. The authors suppose that the lower power of the IRT-LR method may have been caused by the use of larger critical χ^2 values, as was pointed out above. As the number of response options increase, the number of parameters estimated increase for the IRT, but the MACS method. This leads to larger standard errors in the case of the IRT method, especially when sample sizes are small. It would be interesting to see if the MACS performs as well when the τ_{ij} between-category threshold parameters are all estimated. This would make the number of parameters similar across the IRT and CFA frameworks, and the noted disadvantage of the IRT method due to larger critical χ^2 values will no longer hold. However, Stark, et al. also reason that the MACS is a simpler model, and therefore, increase in the number of

response options and violations of normality are less of an issue with the MACS model. Overall, the authors claim that for researchers who are fortunate enough to have a large sample ($N > 1000$) of unidimensional dichotomous nature, IRT procedures may be recommended. However, when scales are factorially complex, and the researcher is interested in the relationships among several latent constructs, CFA-based measures are highly recommended (Stark, Chernyshenko, & Drasgow, 2006).

Gonzalez-Roma, Hernandez, and Gomez-Benito (2006) assessed the power and Type-I error rate of the MACS model in detecting DIF for a unidimensional GRM (5 categories). They simulated data based on true values from analyzing responses for a ten-item 'Team Climate Inventory'. The following variables were manipulated in this study: sample size (100, 200, 400, and 800); sample size ratio (4 conditions of equal sample size, and 6 conditions of unequal sample size, $RG > FG$); latent trait distribution (equal vs. unequal, *focal* 1 σ lower); type of DIF (no DIF, uniform DIF, and non-uniform DIF); and magnitude of DIF (low = .10, medium = .25, and high = .50), and computed the power and Type-I error rates for the above simulated conditions. In order to assess power, they computed the *true positives* or proportion of correct identifications, and in order to assess Type-I error rate, they computed the *false positives* or proportion of incorrect DIF identifications (Gonzalez-Roma, Hernandez, & Gomez-Benito, 2006).

For the no-DIF conditions, the number of *false positives* detected was less than the nominal alpha (.005) in all study conditions, when the latent distributions were equal. However, when the latent distributions were unequal, the proportion of *false positives* depended on the equality or inequality of sample sizes between groups. When the sample sizes were equal between groups, then the Type-I error rate was $> .005$. However, when

the sample sizes were unequal, the proportion of *false positives* notably exceeded .005, particularly for non-uniform DIF in the discrimination parameter. However, even in this case, the highest observed proportion of *false positives* was .026.

For the uniform-DIF conditions, they observed that the proportion of *false positives* in the factor loadings and intercepts increased as the magnitude of DIF increased. Furthermore, when the latent distributions were equal across groups, overall the Type-I error rate was not controlled under the nominal alpha of .005, and this was especially the case when magnitude of DIF and sample sizes increased. However, the highest observed *false positives* rate was .06 in a condition with high DIF magnitude, and where the sample sizes were unequal between groups. However, when the latent distributions were unequal across groups, the overall Type-I error rate was further inflated with error rates close to .10 or higher in several conditions, especially when the DIF magnitude was large, and the sample size ratio was unequal (Gonzalez-Roma, Hernandez, & Gomez-Benito, 2006).

The power of DIF detection (or proportion of *true positives*) increased as DIF magnitude and sample size increased. When DIF magnitude was high, the DIF items were always detected (power = 1), regardless of sample size. However, even with medium DIF magnitude, the DIF detection power was near perfect for sample sizes larger than 200, as long as both groups had equal samples. DIF detection in *true* DIF items was worst in a condition where the sample size ratio was 800:100 and the DIF magnitude was low. However, in general, the conditions where sample size was 100 did not have high power in DIF detection, unless the DIF magnitude was large (Gonzalez-Roma, Hernandez, & Gomez-Benito, 2006). Overall, however, Gonzales-Roma et al. found that

the MACS method had acceptable power ($\geq .70$) in detecting DIF even for conditions with small samples and medium levels of DIF. Furthermore, they found that the power in DIF detection under the MACS model increased as sample size and DIF magnitude increased. Finally, they found that the Type-I error rate was consistently controlled ($<.10$ in all cases) under the MACS model (Gonzalez-Roma, Hernandez, & Gomez-Benito, 2006).

All the authors who used the MG-CFA approach for DIF detection (Gonzalez-Roma, Hernandez, & Gomez-Benito, 2006; Stark, Chernyshenko, & Drasgow, 2006) point out the advantages of the CFA-based approaches for multidimensional models. However, Wu and Lei (2009) were the first to use the MG-CFA approach to detect DIF for a multidimensional dichotomous model. They compared the power and Type-I error of the MG-CFA model in detecting DIF when the model was misspecified as unidimensional to when it was correctly specified as two-dimensional. They point out the importance of testing for invariance in the residual variance in addition to testing for metric and scalar invariance. Increased residual variances give rise to a flatter conditional probability curve, and attenuate the relationship between the latent response variables and the latent traits (Muthen & Asparouhov, 2002). Therefore, they recommend a two-step procedure to detecting DIF from the MG-CFA perspective: first, they recommend testing for invariance in the τ and λ parameters; and if found invariant, they recommend that one also test for invariance in the residual variance parameter, as a second step (Wu & Lei, 2009).

Wu and Lei (2009) used the M2PL model to generate data for a 40-item test with four DIF items. The following variables were manipulated: ability distribution, where

group mean (μ_{RG} and μ_{FG}) difference and correlations between latent traits (ρ) were manipulated for a bivariate distribution ($\mu_{RG}=\mu_{FG}$, $\rho=0$; $\mu_{RG}=\mu_{FG}$, $\rho=.5$; $\mu_{FG}<\mu_{RG}$ by 0.5σ , $\rho=.5$; $\mu_{RG}=\mu_{FG}$, $\rho_{RG}=0.5$, $\rho_{FG}=0$); DIF type (uniform vs. non-uniform); and DIF direction (unidirectional, RG consistently favored; and bi-directional, FG and RG equally favored). They compared a baseline model where all parameters were freely estimated to constrained models where the studied item was constrained to be equal in the subsequent model when studying each item. If the models were significantly different, then the given item was flagged as DIF item, and otherwise, subsequent invariance testing of the error variance parameter was performed for that item (Wu & Lei, 2009).

Wu and Lei (2009) found that Type-I error was greatly inflated (especially in the condition where $\mu_{FG}<\mu_{RG}$ by 0.5σ and $\rho=.5$), when a unidimensional model was used. Additionally, they found that when multidimensionality was not taken into consideration, power was very low, especially for the non-uniform DIF conditions when the bi-directional DIF was in the opposite direction of latent trait differences. However, they found that correctly specifying a multidimensional model tended to significantly reduce Type-I error across all study conditions and power significantly increased even in conditions of non-uniform DIF when the latent trait differences were specified in the opposite direction of bi-directional DIF. Overall, they found that the MG-CFA model was robust in detecting DIF items when the model is correctly specified for latent trait dimensions (Wu & Lei, 2009).

In general, researchers who work with multiple groups and sub-groups have to determine if the instrument is exhibiting DIF for any of those subgroups. They would, overall, be encouraged by this burgeoning of interest in using the MG-CFA approach,

especially when encountered with multidimensional data. As has been pointed out before, a number of assessments in psychology, and in education, are increasingly polytomous in nature (for e.g. non-cognitive and performance-based assessments), and these polytomous models pose different challenges to researchers when compared to dichotomous models. However, no study has investigated the performance of MG-CFA models in assessing DIF for multidimensional polytomous models. Furthermore, Raju's DFIT method that exists for multidimensional dichotomous models has not been extended to multidimensional-polytomous models to assess the effectiveness of a MGRM-based approach to DIF detection. Therefore, assessing and comparing the robustness of MG-CFA based and MIRT based approaches for DIF detection in multidimensional polytomous models is a very important and relevant research question that needs to be addressed.

3.0 METHOD

The two-fold purpose of this study was: (1) to extend Raju's IRT-based DFIT technique to the MGRM and assess its robustness in detecting DIF; and (2) to compare the performance of the MG-CFA based approach to the MGRM-based DFIT approach in detecting DIF items for multidimensional polytomous tests. The mathematical extension of Raju's IRT-based DFIT technique to MGRM was presented in Chapter II. In this chapter, the methodologies used to compare the robustness of the MG-CFA and MGRM-DFIT approach in accurately detecting DIF for multidimensional polytomous items, are presented.

A two-factor 40-item test, with complex structure, and 4 DIF items, was modeled in this study. The test was designed to have 26 multiple-choice, dichotomous items, and 14 performance-based, polytomous items. Both multiple-choice and performance assessment items were modeled in order to mimic real state-level assessments (e.g., PSSA, FCAT, to name a few), which tend to have a mixture of both multiple-choice and performance assessment items. In reality, performance assessment items take a long time to complete. Therefore, in order to improve the breadth of coverage, tests are developed to represent a mix of time-efficient structured items and time-intensive performance items (Messick, 1994).

3.1 DESIGN

A Monte Carlo simulation was performed with one within-subject (estimation method, MG-CFA vs. MGRM-DFIT) independent variable. In addition, five between-subject factors, sample size, sample size ratio, type (Uniform and Non-uniform DIF) of DIF, direction of DIF, and latent mean differences between RG and FG were used in this study (see Table 3.1). These between-subjects independent variables are described subsequently.

The following variables were held constant in the current study, and are described below:

- (i) Proportion of DIF items was held constant in this study, at 10% of the total number of items. This is because previous research (Raju, van der Linder & Fler, 1995; Flowers, Oshima & Raju, 1999) has found that large proportions of DIF (20% or higher) result in very large numbers of *false positives* and *false negatives*. Therefore, subsequent researchers (Bolt, 2002; Gonzalez-Roma, Hernandez & Gomez-Benito, 2006; Stark, Chernyshenko & Drasgow, 2006; Wu & Lei, 2009) have kept the number of DIF items constant, in general, and typically at around 10% (Wu & Lei, 2009) in their studies. Therefore, 4 out of 40 (i.e., 10%) DIF items were modeled in this study. However, all 4 DIF item were polytomous, performance assessment items, and DIF was assessed for the 14 polytomous items in the test.

Table 3-1 Study Design

<u>MANIPULATED VARIABLES</u>	<u>MANIPULATED LEVELS</u>					
<i>I. Sample Size</i>	1000			2000		
<i>II. Sample size Ratio</i>	RG=80% FG=20%		RG=70% FG=30%		RG=50% FG=50%	
<i>III & IV. Type & Direction of DIF</i>	<i>III. Uniform DIF</i> (DIF varied in the intercept ‘b’ parameter)			<i>IV. Non-Uniform DIF</i> (DIF varied in the discrimination ‘a’ parameter)		
	Direction: FG scores less than RG					
	DIF only in the largest ‘b’ parameter across groups (FG less likely to score in the largest category)	DIF in the two largest ‘b’ parameters across groups (FG less likely to score in the two largest categories)		One dimension	Both dimensions	
	Direction: FG scores greater than RG			‘a’ parameter is consistently less discriminating for the FG in one dimension	Same direction (less discriminating for the FG in both dimensions)	Opposite direction (more discriminating for the FG in one dimension and less discriminating for the FG in the other dimension)
	DIF only in the largest ‘b’ parameter across groups (FG more likely to score in the largest category)	DIF in the two largest ‘b’ parameters across groups (FG more likely to score in the two largest categories)				
<i>VI. Distributional differences (Impact)</i>	$\theta_R = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right)$ and $\theta_F = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right)$	$\theta_R = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right)$ and $\theta_F = N\left(\begin{pmatrix} 0 \\ -.5 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right)$		$\theta_R = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right)$ and $\theta_F = N\left(\begin{pmatrix} -.5 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right)$	$\theta_R = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right)$ and $\theta_F = N\left(\begin{pmatrix} -.5 \\ -.5 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right)$	
<i>VII. Estimation method</i>	MG-CFA			MGRM-DFIT		
<u>OUTCOME VARIABLES</u>	<i>Type-I error rate</i> (Proportion of false positives)			<i>Power</i> (Proportion of true positives)		

NOTE:

1. Data generated on 26 dichotomous items and 14 polytomous items. Number of scale-points constant at 5-points for the polytomous items.
2. DIF magnitude held constant for the 'a' parameter at 0.2, and for the 'b' parameter at 0.5.
3. Proportion of DIF is held constant at 10%.
4. Correlation between latent dimensions is held constant at 0.6.

- (ii) The correlation between the latent dimensions was held constant at 0.6. The latent constructs being measured (e.g., mathematical reasoning and mathematical communication) are likely to be related to the same degree, and not vary. Furthermore, most multidimensional latent constructs of interest in educational assessment, are likely to be highly correlated with each other. Therefore, a high dimensional correlation of 0.6 was used in this study.
- (iii) Magnitude of DIF was held constant at 0.5 for the 'β' parameter. In general, researchers (Flowers, Oshima & Raju, 1999; Meade, Lautenschlager & Johnson, 2006; Stark, Chernyshenko & Drasgow, 2006) have repeatedly found that items with a large amounts of DIF are more easily detected than items with small amounts of DIF. Additionally, studies using IRT-based methods (Flowers, Oshima & Raju, 1999; Meade, Lautenschlager & Johnson, 2006) have reported that only larger DIF magnitudes (of .40 or higher) result in high DIF detection power. Therefore, a DIF magnitude of 0.5 was used for the 'β' parameter.
- (iv) Magnitude of DIF had to be held constant at 0.2 for the 'α' parameter. The element $(\sum_{h=1}^H \lambda_{ih}^2 + 2\lambda_{i1}\phi\lambda_{i2})$ in the denominator for computing the population 'α' and 'β' parameters in Eq. (9) and (10) cannot be greater than one (Kannan & Kim, 2009). If it is greater than one, then we would have to take the square root of a negative number, and the denominator would not be estimable (Kannan & Kim, 2009). The correlation between the two latent dimensions was held constant at 0.6 in this study. The 'α' parameter was modeled from a uniform $U(0,1)$ distribution for the dominant dimension, and 0.75 was added to this value. For the secondary (minor) dimension, the 'α' parameter was modeled from a normal $N(0,0.1)$ distribution, and the absolute

value of the generated random number was taken. Therefore, the ' α ' values could range from 0.75 – 1.75 for the dominant dimension, and from 0 – 0.3 for the secondary dimension. Due to the large ' α_1 ' and ' ρ ' values, the values for the ' α_2 ', and the DIF for the ' α ' parameter had to be limited to 0.2, such that Eq. (9) and (10) would be estimable (Kannan & Kim, 2009).

- (v) Various kinds of polytomous scales are frequently used in state and national assessments. While 3-point partial-credit assessment items are common in the National Assessment of Educational Progress (NAEP), state assessments typically use 5-point partial-credit and performance assessment items. However, most of the previous studies using polytomous items (Gonzalez-Roma, Hernandez & Gomez-Benito, 2006; Meade & Lautenschlager, 2004a; Stark, Chernyshenko & Drasgow, 2006) have assessed only 5-point scales. Therefore, in the interest of time, only 5-point polytomous scales were used in this study.

The between-subjects factors that were manipulated are as follows:

- (i) Sample size was varied at two levels (1000, and 2000).
- (ii) Sample-size ratio was varied at three levels (RG = 80%, FG = 20%; RG = 70%, FG = 30%; and RG = 50%, and FG = 50%).
- (iii) Type of DIF was varied at two levels – uniform DIF and non-uniform DIF (see Table 3.1). Further, uniform DIF was varied at two levels: DIF in only the largest 'b' parameter across groups and DIF in the two largest 'b' parameters across groups. Non-uniform DIF (in the 'a' parameter) is varied at two levels: non-uniform DIF in one dimension and non-uniform DIF in both dimensions.

- (iv) Direction of DIF was modeled within Uniform DIF, such that, either the *reference group* performs better than the *focal group* or vice versa. Non-uniform DIF in both dimensions could either be in the same direction or in opposite directions.
- (v) The bivariate latent distribution was varied between the two groups. As described above, the correlation between the latent dimensions was held constant. However, latent mean differences between the RG and FG are varied at four levels (see Table 3.1), where either the FG and RG have equal means on both dimensions, or the FG has a lower mean on one or both dimensions.

The levels of the between subject factors chosen are clarified below, and the rationale for choosing the levels are provided. The between-subject factors were chosen to reflect some of the most frequently used variables in DIF research. Furthermore, the levels at which these factors were manipulated reflect either situations most frequently encountered or situations most frequently used by researchers (Gonzalez-Roma, Hernandez & Gomez-Benito, 2006; Meade & Lautenschlager, 2004a; Raju, Laffitte & Byrne, 2002; Stark, Chernyshenko & Drasgow, 2006; Wu & Lei, 2009).

Large samples are typically encountered in applied research within the field of educational testing. However, a number of previous studies using IRT-based DIF methods (Bolt, 2002; Meade & Lautenschlager, 2004a; Meade, Lautenschlager & Johnson, 2006; Raju, van der Linden & Fleer, 1995, to name a few) found that conditions with smaller sample sizes, especially for polytomous items (Bolt, 2002; Meade & Lautenschlager, 2004a), gave rise to the most inconsistencies in DIF detection (higher *false positives* and *false negatives*). On the other hand, studies using MG-CFA based measures (Gonzalez-Roma, Hernandez & Gomez-Benito, 2006; Stark, Chernyshenko &

Drasgow, 2006) found that these measures produced stable and consistent results at smaller sample sizes. Therefore, sample sizes of (i) 1000 and (ii) 2000 were used in this study.

Different types of *reference* and *focal groups* are typically encountered in applied research, and the sample-sizes within these sub-groups are found in varying proportions. For some sub-groups, such as those that are gender-based, the sample-size ratio for the two groups is more likely to be equal (50-50). However, for other sub-groups of interest, such as those based on racial and ethnic minorities, the sample-size ratio encountered for *reference* and *focal groups* are more likely to be disparate (70-30 or 80-20). Furthermore, previous research (Gonzalez-Roma, Hernandez & Gomez-Benito, 2006) has found that unequal sample sizes result in larger Type-I error rates in DIF detection.

Therefore, three different sample-size ratios were used in this study:

- (i) 80% of the sample comes from the RG, and 20% of the sample comes from the FG
- (ii) 70% of the sample comes from the RG, and 30% of the sample comes from the FG
- (iii) 50% of the sample comes from the RG, and 50% of the sample comes from the FG

One of the most frequently assessed factors in DIF research is type (*uniform* and *non-uniform*) of DIF (Gonzalez-Roma, Hernandez & Gomez-Benito, 2006; Meade & Lautenschlager, 2004a; Meade, Lautenschlager & Johnson, 2006; Raju, van der Linden & Fler, 1995; Stark, Chernyshenko & Drasgow, 2006; Wu & Lei, 2009). Within the IRT framework, *uniform* and *non-uniform* DIF are used with *reference* to differences in the intercept (β) and discrimination (α) parameter across groups. When the groups differ only in the intercept or item difficulty parameter, then this is called *uniform* DIF (Gonzalez-Roma, Hernandez & Gomez-Benito, 2006). However, when there are

differences in the item discrimination parameter, this is referred to as *non-uniform* DIF (Gonzalez-Roma, Hernandez & Gomez-Benito, 2006).

While some researchers (Meade & Lautenschlager, 2004a) have found that IRT-based methods have higher power for detecting *uniform* DIF in polytomous items, other researchers (Stark, Chernyshenko & Drasgow, 2006) have found that the CFA-based methods have higher power for detecting *uniform* DIF, especially for polytomous items. In addition, some researchers (Gonzalez-Roma, Hernandez & Gomez-Benito, 2006; Stark, Chernyshenko & Drasgow, 2006) have found that power for detecting *non-uniform* DIF was significantly related to sample size. Therefore, *uniform* and *non-uniform* types of DIF were manipulated in this study.

Furthermore, different types of *uniform* and *non-uniform* DIF have been assessed in previous research. DIF would exist within different ' β ' parameters, when the *reference group* (RG) is more likely to score in certain score categories when compared to the *focal group* (FG), or vice versa. Three types of *uniform* DIF have been typically assessed in previous research (Meade & Lautenschlager, 2004a; Meade, Lautenschlager & Johnson, 2006): DIF in the highest score category, DIF in the two highest score categories, and DIF in the two extreme score categories.

[NOTE: The third type of *uniform DIF* used in previous research is more relevant to cross-cultural psychological survey research where central tendency in survey responses is often found with certain subgroups and cultures (Hofstede, 1984; 2001). Central tendency in survey responses refers to a tendency to stick to the middle response categories in rating scales and surveys, and avoid extreme responses. The FG might be less likely to respond in the extreme score categories in any scale. However, psychological scales are not relevant within the context of the complex-structure multidimensional model used in this study. Therefore, this last type of *uniform DIF* was not modeled here and only the first two types of *uniform DIF* were manipulated (see Table 3.1).]

DIF would exist in the highest score category if the RG (or the FG) has a higher probability of scoring in the highest score category (β parameter). For the 5-point scales, this would refer to the score categories of (5). Additionally, the FG (or the RG) might be

less likely to score not only in the highest score category, but in a couple of the highest score categories, and this would present DIF for the two (or more) highest score categories. Therefore, DIF in two of the highest score categories is also modeled here. For the 5-point scale, this would correspond to the score categories of (4&5). In order to manipulate DIF direction for *uniform DIF*, in some conditions, the *focal group* (FG) was manipulated to have a lower probability of scoring in category 4 and 5, and in other conditions, the *focal group* (FG) was manipulated to have a higher probability.

Therefore, *uniform DIF* was manipulated at four levels:

- (i) the FG has a lower probability of scoring a 5
- (ii) the FG has a lower probability of scoring a 4 & 5
- (iii) the FG has a higher probability of scoring a 5
- (iv) the FG has a higher probability of scoring a 4 & 5

Non-uniform DIF is typically assessed in either one or both dimensions for the discrimination parameter (Wu & Lei, 2009). When *non-uniform DIF* is modeled in one dimension, the ' α ' parameter for the one latent dimension is varied for the DIF items, such that the item is more discriminating for the RG (or FG) in that dimension. When *non-uniform DIF* is modeled in both dimensions, the ' α ' parameter was varied for the DIF items on both the latent dimensions. A number of studies (Raju, van der Linden & Fler, 1995; Flowers, Oshima & Raju, 1999; Wu & Lei, 2009) also manipulate DIF direction (*unidirectional* or *balanced bi-directional*) for *non-uniform DIF* in both dimensions. For *non-uniform DIF* in both dimensions, unidirectional and bi-directional DIF was modeled. If the *non-uniform DIF* is in the same direction for both dimensions, then the item would be less discriminating for the FG in both dimensions. However, if the

non-uniform DIF is in opposite directions, then the item would be less discriminating for the FG in one dimension, and less discriminating for the RG in the other dimension (see Table 3.1).

Therefore, *non-uniform* DIF was varied at five levels:

- (i) *non-uniform* DIF in one dimension: item is less discriminating for the FG only on the second dimension
- (ii) *non-uniform* DIF in one dimension: item is less discriminating for the FG only on the first dimension
- (iii) *unidirectional non-uniform* DIF: item is less discriminating for the FG on both dimensions
- (iv) *bidirectional non-uniform* DIF: item is less discriminating for the FG in the first dimension and more discriminating for the FG in the second dimension
- (v) *bidirectional non-uniform* DIF: item is more discriminating for the FG in the first dimension and less discriminating for the FG in the second dimension.

Lastly, unequal latent distributions between the *reference* and *focal groups* have been found to inflate Type-I error rate (Gonzalez-Roma, Hernandez & Gomez-Benito, 2006; Wu & Lei, 2009). Furthermore, the power of DIF detection in the presence of *impact* has long been interesting to researchers (Bolt, 2002; Flowers, Oshima & Raju, 1999; Stark, Chernyshenko & Drasgow, 2006). Since a multidimensional model was simulated here, data was generated from a bivariate normal distribution. The correlation between the latent dimensions was held constant. Therefore, latent mean differences between the RG and FG were varied at four levels resulting in the following four conditions:

- (i) the μ for the *reference group* and *focal group* are equal on both latent dimensions:

$$\theta_R = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right) \text{ and } \theta_F = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right);$$

- (ii) *focal group* has lower μ on the second dimension, but not on the first dimension:

$$\theta_R = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right) \text{ and } \theta_F = N\left(\begin{pmatrix} 0 \\ -0.5 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right);$$

- (iii) the *focal group* has a lower μ on the first dimension, but not on the second dimension:

$$\theta_R = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right) \text{ and } \theta_F = N\left(\begin{pmatrix} -0.5 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right);$$

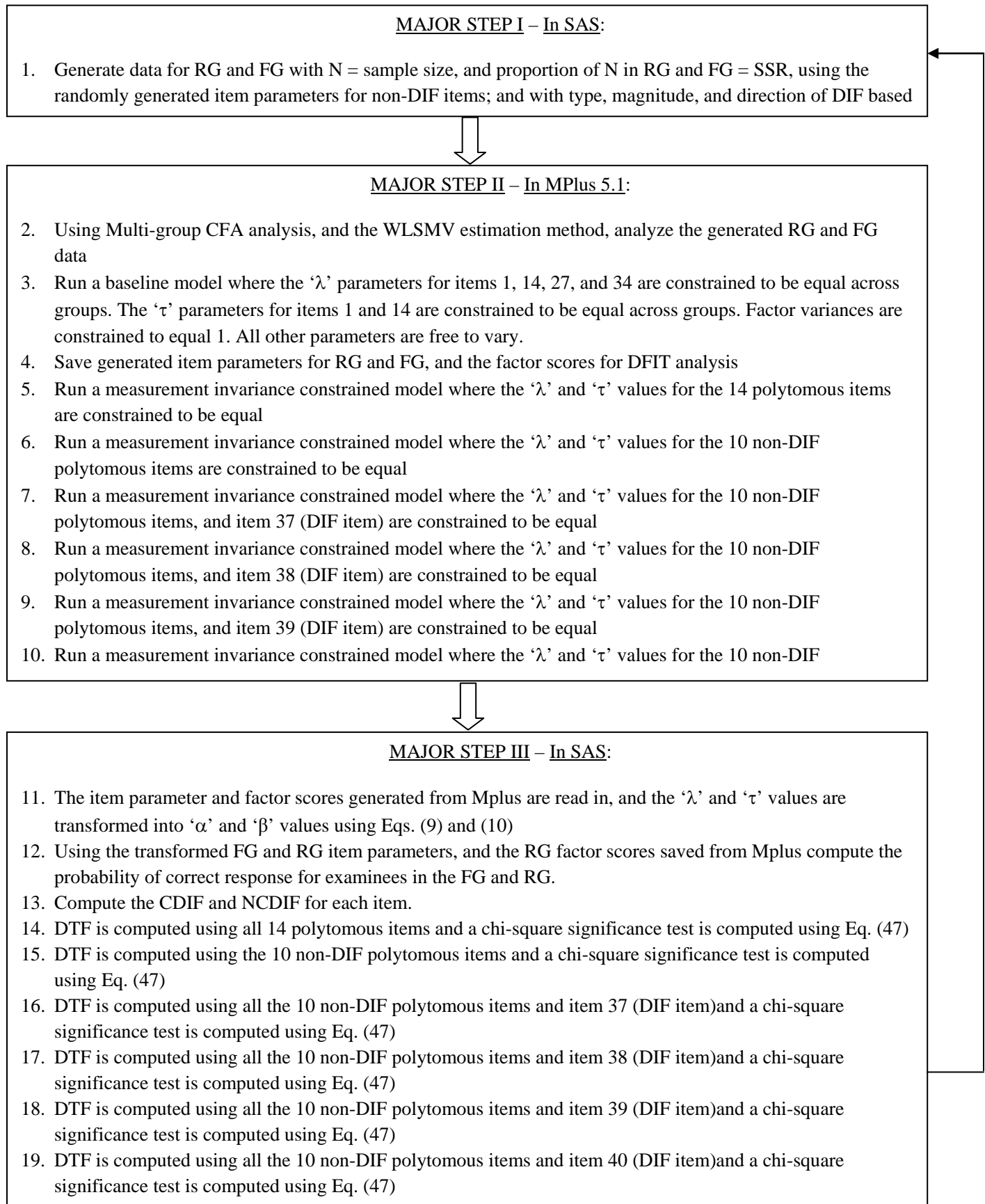
- (iv) the *focal group* has a lower μ on both latent dimensions:

$$\theta_R = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right) \text{ and } \theta_F = N\left(\begin{pmatrix} -0.5 \\ -0.5 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}\right).$$

3.2 DATA GENERATION

The data generation scheme is described in this section, and the simulation flow chart is presented in Table 3.2 below. The data was generated from a two-factor 40-item MIRT model. The data for the 26 dichotomous items were generated from a 2PL MIRT model, and the data for the 14 polytomous items were generated from the MGRM. Items 1, 14, 27, and 34 were unidimensional items, where items 1 and 27 loaded only on the first factor and items 14 and 34 loaded only on the second factor. All the remaining items were multidimensional. However, items 1-13, 27-33 loaded predominantly on the first dimension, and items 14-26, 34-40 loaded predominantly on the second dimension.

Table 3-2 Simulation Flow Chart



Two separate datasets were generated for the *reference group* and the *focal group* samples. Eighty two replications were performed within each cell (2x3x4x5x4x2=total of 960 cells) in the design. For each replication, for each condition, the data simulation steps are presented as a flowchart and described below.

The SAS (Statistical Analysis Software) program was used to simulate subject responses to the 40 items. Latent ability estimates were first generated for ‘N’ subjects’ from a standard multivariate normal distribution $N\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} 1 & \phi \\ \phi & 1 \end{pmatrix}\right)$. The ‘N’ depended on the sample size condition, and the μ (mean) for the *focal group* (FG) depended on the latent distribution condition (see Table 3.1). The ‘ α ’ parameter was modeled from a uniform $U(0,1)$ distribution for the dominant dimension, and 0.75 was added to this value. For the secondary (minor) dimension, the ‘ α ’ parameter was modeled from a normal $N(0,0.1)$ distribution, and the absolute value of the generated random number was taken. Therefore, the ‘ α ’ values could range from 0.75 – 1.75 for the dominant dimension, and from 0 – 0.3 for the secondary dimension.

The low ‘ α_2 ’ values modeled do not reflect a truly multidimensional model, these values reflect a secondary factor loading of $< \sim .15$. The loading on the secondary ‘ α_2 ’ had to be limited due to the computational limitations of Eq. (9) and (10) mentioned in the study design. The element $(\sum_{h=1}^H \lambda_{ih}^2 + 2\lambda_{i1}\phi\lambda_{i2})$ in the denominator for computing the population ‘ α ’ and ‘ β ’ parameters in Eq. (9) and (10) cannot be greater than one (Kannan & Kim, 2009). If it is greater than one, then we would have to take the square root of a negative number, and the denominator would not be estimable (Kannan & Kim, 2009).

The ‘ β ’ threshold parameter for the 26 dichotomous items was generated from a $N(0,0.25)$ distribution. For the 14 polytomous items, the ‘ β ’ parameter for the first category was generated from a $N(-0.5,0.25)$ distribution. Threshold values were then generated from a uniform $U(0,1)$ distribution, and the square-root of this value was added to the generated ‘ β ’ parameter for the first category to generate ‘ β ’ parameter for the remaining category thresholds, respectively. This enabled a uniform distance between the response category thresholds, and therefore assured an underlying graded model.

A DIF of 0.2 was introduced in one or both of the ‘ α ’ parameters, and a DIF of 0.5 was introduced in either the highest or two highest ‘ β ’ parameters. Category response functions (CRFs) for a 5-point scale demonstrating such DIF are presented in Figures 3.1 through 3.5. Figure 3.1 represents the category response functions for each of five categories for the *focal* and *reference groups* when no DIF is introduced. The same item parameters ($\alpha_1 = 0.92$, $\alpha_2 = 0.20$, $\beta_1 = -0.79$, $\beta_2 = -0.01$, $\beta_3 = 0.76$, $\beta_4 = 1.54$) are used to generate these CRFs for both the *reference group* and the *focal group* in this figure. It is clear from the CRFs that the probability of scoring in each category does not differ for the *reference* and *focal groups* here.

Figure 3.2 represents the category response functions for each of five categories for the *focal* and *reference groups* when a DIF of 0.5 is introduced in the last ‘ β ’ parameter ($\alpha_1 = 1.13$, $\alpha_2 = 0.19$, $\beta_1 = -0.88$, $\beta_2 = -0.05$, $\beta_{3_rg} = 0.78$, $\beta_{4_rg} = 1.60$, $\beta_{3_fg} = 0.78$, $\beta_{4_fg} = 2.10$). No DIF has been introduced in the ‘ α ’ parameter in generating these CRFs. It can be seen from Figure 3.2 that the probability of scoring in the first three categories does not differ for the *reference* and *focal groups* when DIF is introduced only in the last ‘ β ’ parameter. However, when a DIF of 0.5 is introduced in the last ‘ β ’

parameter, it becomes more difficult for the *focal group* to score in the highest score category. Therefore, the probability of scoring a '4' becomes higher for the *focal group* than the *reference group*, and the probability of scoring a '5' becomes higher for the *reference group* than the *focal group*.

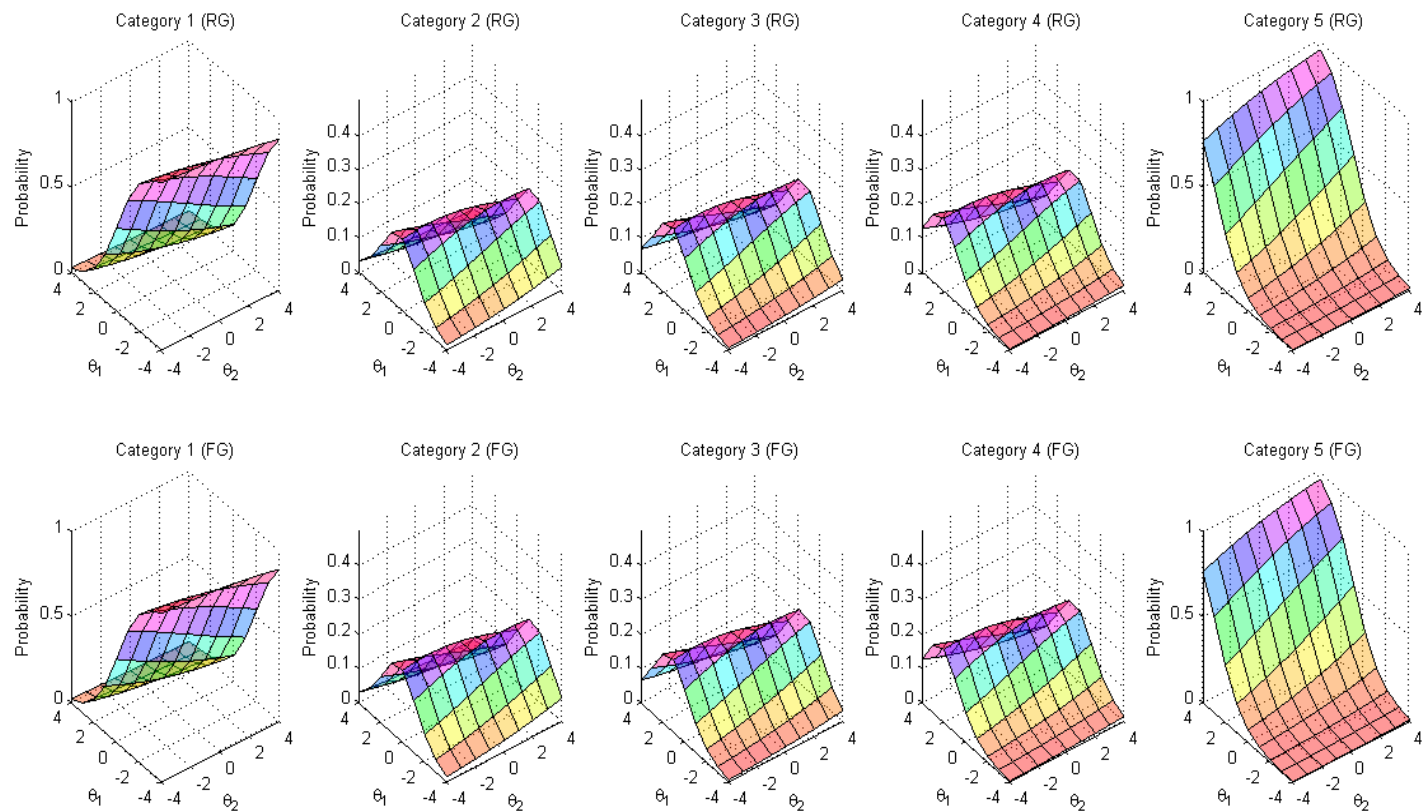


Figure 3-1 Category Response Functions for a 5-point item for the *reference* and *focal group* when no DIF is introduced.

$[\alpha_1 = 0.92, \alpha_2 = 0.20, \beta_1 = -0.79, \beta_2 = -0.01, \beta_3 = 0.76, \beta_4 = 1.54]$

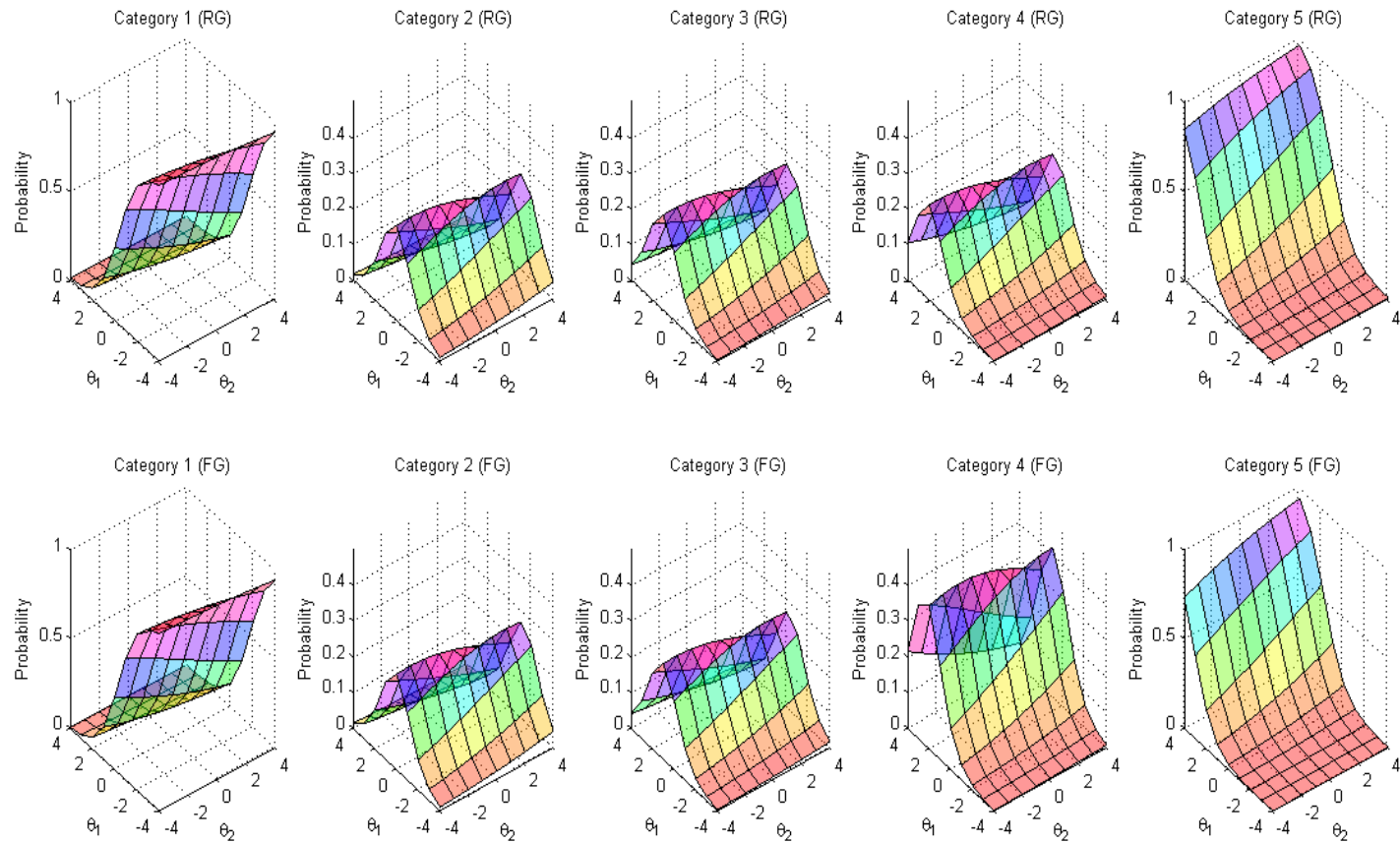


Figure 3-2 Category Response Functions for a 5-point item for the *reference* and *focal group* when no DIF is introduced in the 'a' parameter, and DIF of 0.5 is introduced in the highest 'b' parameter.

$[\alpha_1 = 1.13, \alpha_2 = 0.19,$
 $\beta_1 = -0.88, \beta_2 = -0.05, \beta_{3_rg} = 0.78, \beta_{4_rg} = 1.60, \beta_{3_fg} = 0.78, \beta_{4_fg} = 2.10]$

Figure 3.3 represents the category response functions for each of five categories for the *focal* and *reference groups* when a DIF of 0.5 is introduced in the last two ‘ β ’ parameters ($\alpha_1 = 1.56$, $\alpha_2 = 0.23$, $\beta_1 = -0.42$, $\beta_2 = 0.24$, $\beta_{3_rg} = 0.91$, $\beta_{4_rg} = 1.57$, $\beta_{3_fg} = 1.41$, $\beta_{4_fg} = 2.07$). No DIF has been introduced in the ‘ α ’ parameter in generating these CRFs. It can be seen from Figure 3.3 that the probability of scoring in the first two categories does not differ for the *reference* and *focal groups* when DIF is introduced in the last two ‘ β ’ parameters. However, when a DIF of 0.5 is introduced in the both highest ‘ β ’ parameters, it becomes more difficult for the *focal group* to score in the two highest score category. Therefore, the probability of scoring a ‘3’ becomes higher for the *focal group* than the *reference group*, and the probability of scoring a ‘4’ and a ‘5’ becomes higher for the *reference group* than the *focal group*.

Figure 3.4 represents the category response functions for each of five categories for the *focal* and *reference groups* when a DIF of 0.2 is introduced in both ‘ α ’ parameters, and a DIF of 0.5 is introduced in the last ‘ β ’ parameters ($\alpha_{1_rg} = 0.98$, $\alpha_{2_rg} = 0.23$, $\alpha_{1_fg} = 1.18$, $\alpha_{2_fg} = 0.43$, $\beta_1 = -0.92$, $\beta_2 = -0.47$, $\beta_{3_rg} = -0.03$, $\beta_{4_rg} = 0.42$, $\beta_{3_fg} = -0.03$, $\beta_{4_fg} = 0.92$). It can be seen from Figure 3.4 that the probability of scoring in the first three categories does not differ for the *reference* and *focal groups* when DIF is introduced only in the last ‘ β ’ parameters. However, the items become more discriminating for the *focal group* when a DIF of 0.2 is introduced in the ‘ α ’ parameter. Furthermore, when a DIF of 0.5 is introduced in the both highest ‘ β ’ parameters, it becomes more difficult for the *focal group* to score in the highest score category. Therefore, the probability of scoring a ‘4’ becomes higher for the *focal group* than the

reference group, and the probability of scoring a ‘5’ becomes higher for the *reference group* than the *focal group*.

Figure 3.5 represents the category response functions for each of five categories for the *focal* and *reference groups* when a DIF of 0.2 is introduced in both ‘ α ’ parameters, and a DIF of 0.5 is introduced in the last two ‘ β ’ parameters ($\alpha1_{rg} = 1.49$, $\alpha2_{rg} = 0.28$, $\alpha1_{fg} = 1.69$, $\alpha2_{fg} = 0.48$, $\beta1 = -0.20$, $\beta2 = 0.15$, $\beta3_{rg} = 0.50$, $\beta4_{rg} = 0.86$, $\beta3_{fg} = 1.00$, $\beta4_{fg} = 1.36$). It can be seen from Figure 3.5 that the probability of scoring in the first two categories does not differ for the *reference* and *focal groups* when DIF is introduced in the last two ‘ β ’ parameters. However, the items become more discriminating for the *focal group* when a DIF of 0.2 is introduced in the ‘ α ’ parameter. Furthermore, when a DIF of 0.5 is introduced in the both highest ‘ β ’ parameters, it becomes more difficult for the *focal group* to score in the two highest score category. Therefore, the probability of scoring a ‘3’ becomes higher for the *focal group* than the *reference group*, and the probability of scoring a ‘4’ and a ‘5’ becomes higher for the *reference group* than the *focal group*.

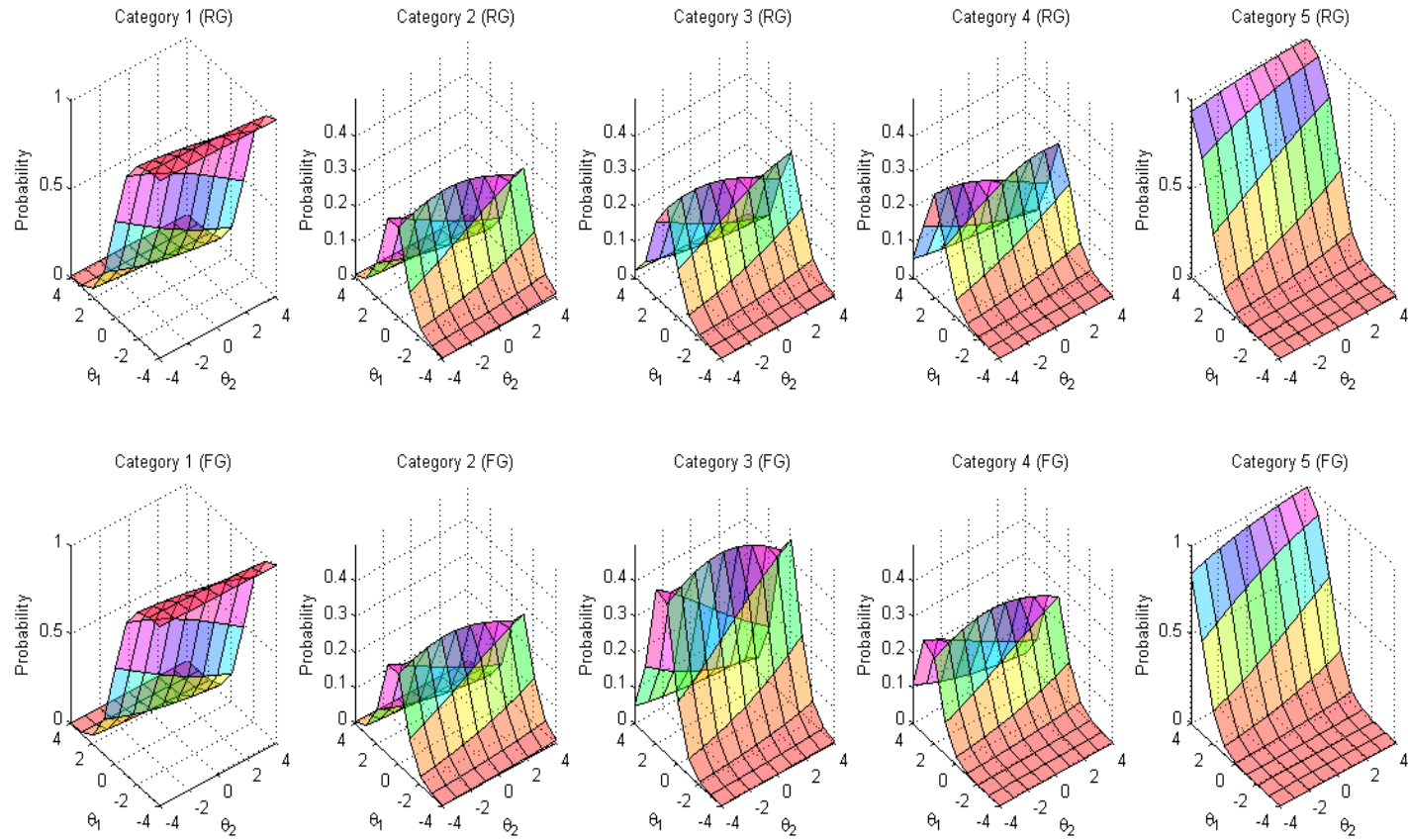


Figure 3-3 Category Response Functions for a 5-point item for the *reference* and *focal* group when no DIF is introduced in the 'a' parameter, and DIF of 0.5 is introduced in the two highest 'b' parameter.

[$\alpha_1 = 1.56, \alpha_2 = 0.23,$
 $\beta_1 = -0.42, \beta_2 = 0.24, \beta_3_{rg} = 0.91, \beta_4_{rg} = 1.57, \beta_3_{fg} = 1.40, \beta_4_{fg} = 2.07]$

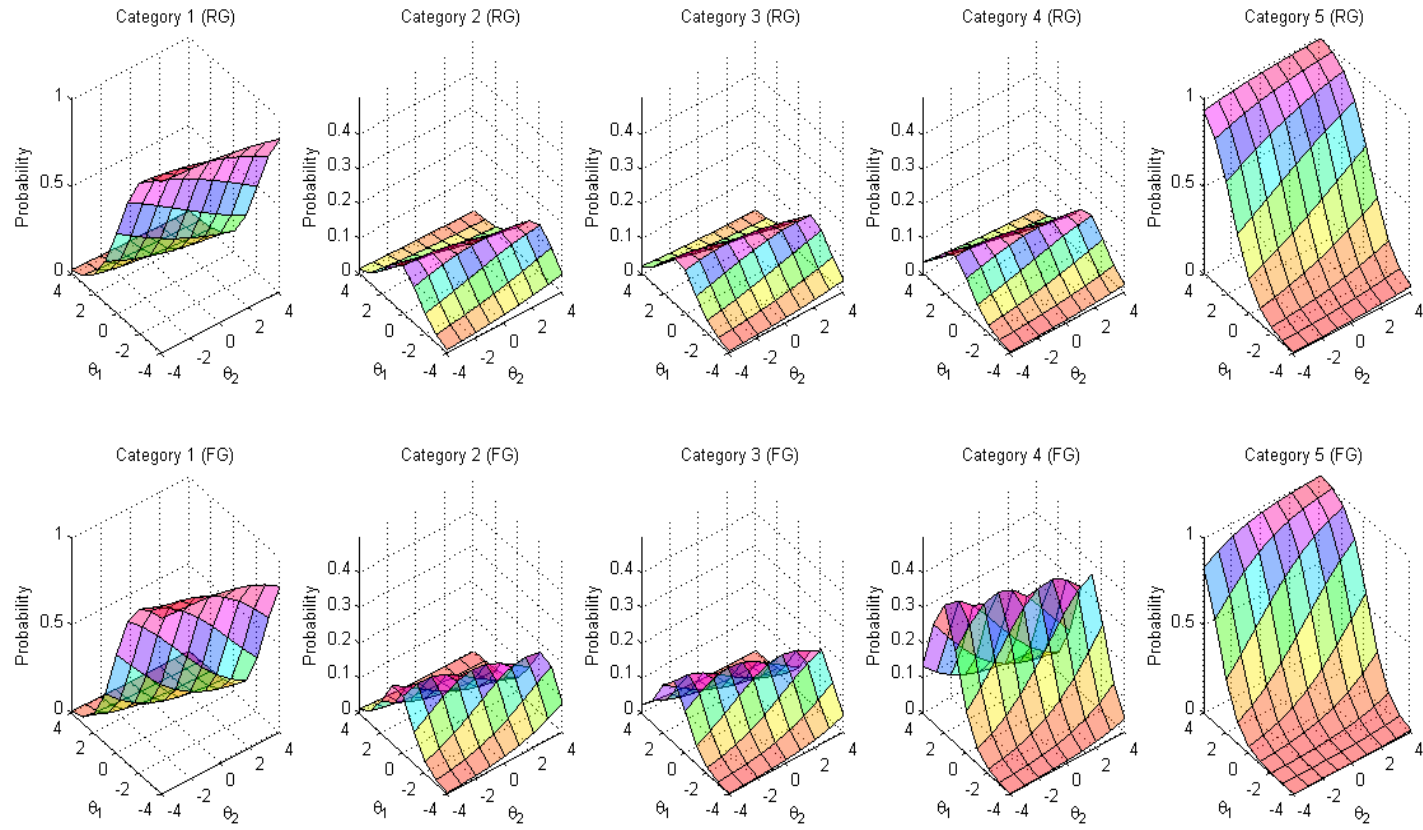


Figure 3-4 Category Response Functions for a 5-point item for the reference and focal group when a DIF of 0.2 is introduced in both 'a' parameters, and DIF of 0.5 is introduced in the highest 'b' parameter.

$[\alpha_{1_rg} = 0.98, \alpha_{2_rg} = 0.23, \alpha_{1_fg} = 1.18, \alpha_{2_fg} = 0.43,$
 $\beta_1 = -0.92, \beta_2 = -0.47, \beta_{3_rg} = -0.03, \beta_{4_rg} = 0.42, \beta_{3_fg} = -0.03, \beta_{4_fg} = 0.92]$

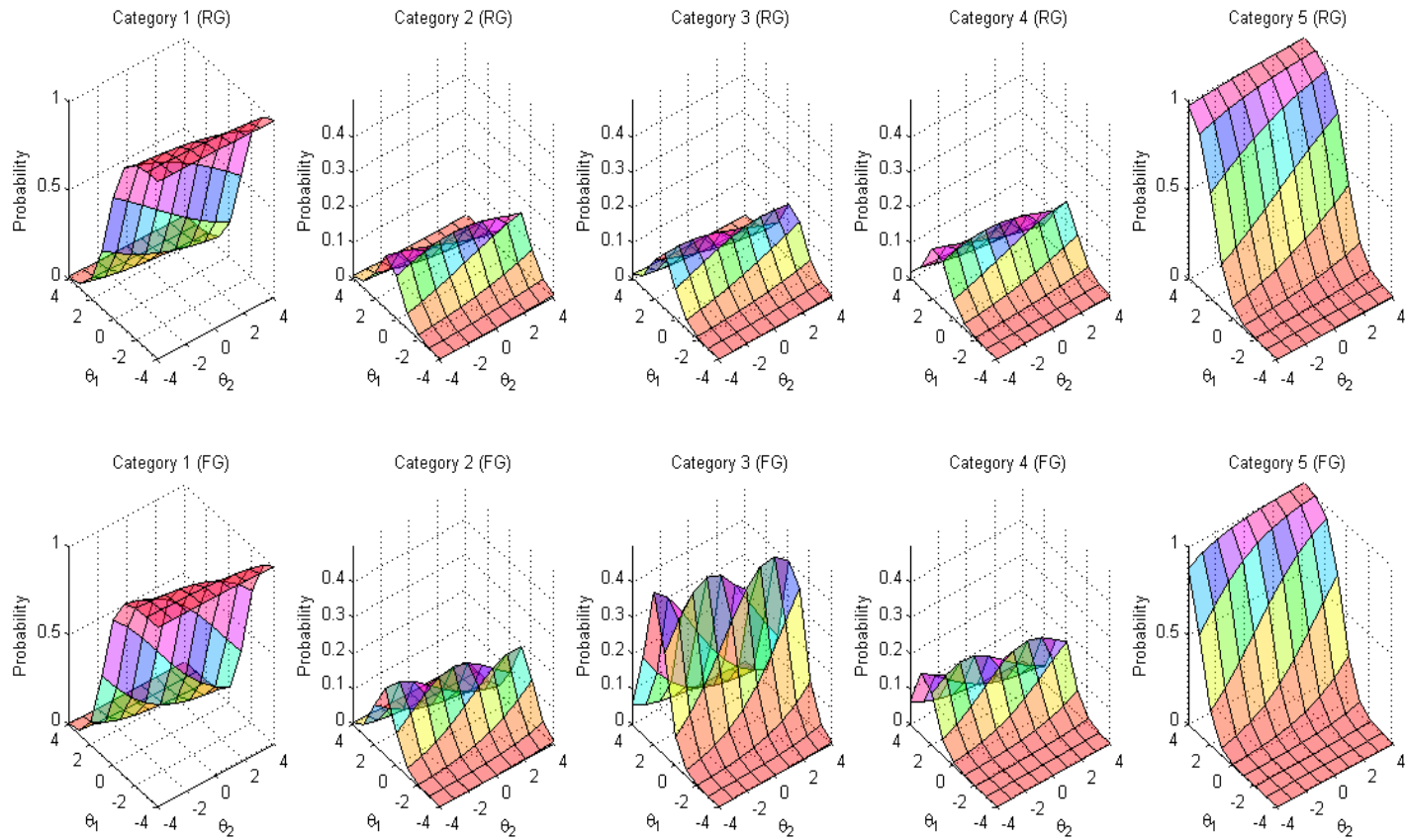


Figure 3-5 Category Response Functions for a 5-point item for the reference and focal group when a DIF of 0.2 is introduced in both 'a' parameters, and DIF of 0.5 is introduced in the two highest 'b' parameter.

$[\alpha1_{rg} = 1.49, \alpha2_{rg} = 0.28, \alpha1_{fg} = 1.69, \alpha2_{fg} = 0.48,$
 $\beta1 = -0.20, \beta2 = 0.15, \beta3_{rg} = 0.50, \beta4_{rg} = 0.86, \beta3_{fg} = 1.00, \beta4_{fg} = 1.36]$

Using these generated item parameters, the probability of correct response was calculated using Eq. (33) for the 26 dichotomous items, and using Eq. (3) for the 14 polytomous items. To obtain discrete scores, the calculated response was compared to a randomly generated number from a uniform $U(0,1)$ distribution. The student's response is 1 if the probability is greater than the random number and 0 otherwise. Discrete data was read into Mplus 5.21 (Muthen & Muthen, 2006a; 2006b). Using the MG-CFA analysis, and the WLSMV estimation method, measurement invariance was assessed for the RG and FG item parameters.

A baseline model was first estimated where all parameters are allowed to freely vary across groups. The factor variances were fixed at 1.0, for model identification purposes. Additionally, the ' λ ' parameters for the unidimensional items (items 1, 14, 27 and 34) were constrained to be equal across groups; and the ' τ ' parameters for items 1 and 14 were constrained to be equal across groups. All the remaining item parameters were free to vary across groups. The factor means for the RG were fixed at 0.0, while the factor means for the FG were freely estimated. The factor correlation was freely estimated across groups. Finally, item parameters, correlation between dimensions, and the latent factor scores for the examinees were estimated from the baseline model, and saved for the MGRM-DFIT method.

In the next steps, measurement invariance (constrained) models were analyzed. In addition to the model constraints enforced in the baseline model, parameters for different items were constrained in subsequent models. In the first measurement invariance model, the item parameters for all 14 polytomous items were constrained to be equal across groups. Therefore, in this constrained model, all 14 polytomous items were tested for

DIF. In the second model, item parameters for the 10 non-DIF polytomous items (i.e., items 27-36) were constrained to be equal across groups. This model tests for DIF in the non-DIF items, and was an indicator of the Type-I error rate for the method. In the third, fourth, fifth, and sixth models, item parameters for the 10 non-DIF polytomous items (i.e., items 27-36), and item 37, 38, 39, and 40, respectively, were constrained to be equal across groups. These models test for DIF in the 4 manipulated DIF items.

In other words, the following model comparisons were performed in Mplus:

1. A baseline model where the ' λ ' parameters for items 1, 14, 27, and 34 are constrained to be equal across groups. The ' τ ' parameters for items 1 and 14 are constrained to be equal across groups. Factor variances are constrained to equal 1. All other parameters are free to vary.
2. A measurement invariance constrained model where the ' λ ' and ' τ ' values for the 14 polytomous items are constrained to be equal
3. A measurement invariance constrained model where the ' λ ' and ' τ ' values for the 10 non-DIF polytomous items are constrained to be equal
4. A measurement invariance constrained model where the ' λ ' and ' τ ' values for the 10 non-DIF polytomous items, and item 37 (DIF item) are constrained to be equal
5. A measurement invariance constrained model where the ' λ ' and ' τ ' values for the 10 non-DIF polytomous items, and item 38 (DIF item) are constrained to be equal
6. A measurement invariance constrained model where the ' λ ' and ' τ ' values for the 10 non-DIF polytomous items, and item 39 (DIF item) are constrained to be equal
7. A measurement invariance constrained model where the ' λ ' and ' τ ' values for the 10 non-DIF polytomous items, and item 40 (DIF item) are constrained to be equal

Chi-square difference tests between the constrained models and the baseline model were performed in Mplus and the chi-square values were saved for all six constrained model comparisons. If the difference in chi-square is significant, then the parameters being tested are not invariant across groups, or in other words, DIF would be demonstrated for the constrained parameters. If the difference in chi-square between the constrained and baseline model for a parameter is not significant, then the parameter are invariant across groups.

The Mplus estimates of factor loadings (λ_{ih}) and threshold values (τ_{ij}) for the baseline estimation were read into SAS, and converted into α_{ih} and β_{ij} parameters using Equations (9) and (10). It should be noted that multidimensional linking was not performed for the RG and FG item parameters. Since the linking only performs a linear transformation, the expected scores should not change. From a factor analysis perspective, a linear transformation reflects factor rotation, and therefore, the expected models for the variance-covariance matrix would not differ. Analogously, a linear transformation should not change the expected scores for the DFIT method. In order to test this hypothesis, the DFIT method was employed for one condition with and without linking. The resulting chi-square test results and CDIF & NCDIF values did not differ significantly between linked and unlinked FG parameters. Therefore, the multidimensional linking step was not performed in this simulation.

In addition to item parameters, factor scores were estimated for all examinees under the baseline model using Mplus, and saved. These factor scores for both the RG and the FG were also read into SAS. However, only the RG factor scores were used along with the transformed RG and FG item parameters compute the probability of responding in each category, for each of the 14 polytomous items, for each examinee using Eq. (3). Once the probability of responding in each category was estimated, the *expected item score* (ES_i) was calculated for each item using Eq. (31). In order to estimate the ES_i , the estimated item category scores from Mplus were used. Next, the *expected test response function* (T_s) for each examinee was calculated using Eq. (32). The DTF across the entire group of examinees was estimated using Eq. (25). Finally, the $CDIF_i$ and $NCDIF_i$ for each item were calculated using Equations (29) and (30), respectively.

Six separate DFIT tests were performed each time with different number of items in the total test. First, all 14 polytomous items were used in computing the DTF [Eq. (25)], and the chi-square significance test [Eq. (47)] was performed to test for *Differential Test Functioning*. Second, the 10 non-DIF items were used in computing the DTF, and the chi-square significance test was performed to test for *Differential Test Functioning*. In the subsequent four models, each of the 4 DIF items were added to the 10 non-DIF items, one at a time, and 11 items were used in computing DTF in each of these tests. For each of these tests, the chi-square significance test was performed to test for *Differential Test Functioning*. The chi-square values and ‘p’ values were saved for all of these DFIT tests.

3.3 DATA ANALYSIS

3.3.1 Outcome measures recorded.

Chi-square model comparisons were performed for the following outcome measures: (i) using all 14 polytomous items; (ii) using the 10 non-DIF items; (iii) using the 10 non-DIF items + item 37; (iv) using the 10 non-DIF items + item 38; (v) using the 10 non-DIF items + item 39; and (vi) using the 10 non-DIF items + item 40. The chi-square difference values and ‘p’ values for each of the 6 model comparison tests (see Flowchart) performed under the MG-CFA and the MGRM-DFIT method were recorded for each replication in each condition. Six categorical outcome variables for the each of these tests

were computed by dichotomizing the chi-square ' p ' values as significant at $\alpha = .05$, to indicate DIF / no DIF.

3.3.2 Type-I error rate and Power.

Type-I error rates (or *False Positive* ratio) was computed for each estimation method by tabulating the proportion of replications where the 10 non-DIF items were detected as having significant DIF. Power (or *True Positive* ratio) was computed for each estimation method by tabulating the proportion of replications where significant DIF was detected in the overall test, and the proportion of replications where the 4 DIF items were detected as having significant DIF.

3.3.3 Generalized Estimating Equations (GEE).

A repeated measures logistic regression was performed on the six categorical outcome variables, predicted by estimation method and the between-subjects independent variables (i.e., sample size, SSR, scale-points, uniform DIF, non-uniform DIF, and distributional differences) using generalized estimating equation (GEE). GEE is a marginal model where regression estimates are computed averaged across subjects, while adjusting for the lack of independence in the observations (i.e., repeated measures).

SAS's GENMOD procedure enables GEE analysis by specifying a "repeated" statement in which clustering information and a working correlation matrix are specified.

An “independent” correlation structure was chosen for all analyses, since the correlation among the estimation methods was close to zero. Main effects and higher-order interactions (up to three-way interactions) between estimation method and the other five between-subjects independent variables were examined. Results are interpreted for significant effects with odds ratio that at least represents a 10% effect size (a beta estimate of ± 0.1). Therefore, results are interpreted for significant effects with odds ratio $> .905$ and odds ratio < 1.105 .

4.0 RESULTS

This chapter presents the results obtained from the simulation study. A Monte Carlo simulation was performed with one within-subject factor (estimation method, MG-CFA vs. MGRM-DFIT) and five between-subject factors, sample size, sample size ratio, type (Uniform and Non-uniform DIF) of DIF, direction of DIF, and latent mean differences between RG and FG (see Table 3.1). In total, 82 replications were performed within the 960 conditions ($2 \times 3 \times 4 \times 5 \times 4 \times 2$) in the design. For each condition, in each replication, six chi-square tests (each of 4 DIF items, 10 non-DIF items, and all 14 polytomous items) were performed for each estimation method.

Categorical outcome variables were computed by dichotomizing these chi-square significance tests at $\alpha=.05$. The important research questions addressed in this study were related to the Type-I error rate and power of the two estimation methods. Type-I error rate was computed as the proportion of replications where significant DIF was detected for the 10 non-DIF items. Power was computed as the proportion of replications where significant DIF was detected in the overall test, and for the 4 DIF items.

A repeated measures logistic regression was performed on the six categorical outcome variables, predicted by estimation method and the between-subjects independent variables (i.e., sample size, SSR, scale-points, uniform DIF, non-uniform DIF, and distributional differences) using generalized estimating equation (GEE). Main effects and

higher-order interactions (up to three-way interactions) between estimation method and the other five between-subject independent variables were examined. Several significant main effects and interaction terms were noted. However, the beta parameter values associated with these effects were negligible indicating no practical or interpretive value to these effects. In general, it is not unusual to find several statistically significant effects with low effect sizes in simulation studies (Harwell, Stone, Hsu, & Kirisci, 1996). With greater number of replications, the power of the study tends to be higher. However, there is a difference between significant results and meaningful results. Therefore, results are interpreted only for significant effects with odds ratio that at least represents a 10% effect size, i.e., effects with odds ratio >1.105 and $<.905$ are presented.

The proportion of replications where DIF was detected by each estimation method for each Dependent variable is presented in Tables 4.1 through 4.9. Results from the logistic regression with beta parameter estimates and odds ratio for the statistically significant effects are presented in Appendix-A. The Type-I error rate and Power for the two estimation methods are interpreted in this section. In each of these tables, the proportion of replications where DIF was detected for the 10 non-DIF items indicates the Type-I error rate of the estimation method. The proportion of replications where DIF was detected for each DIF item and for the overall test reflects the Power of the estimation method.

Results are presented for the within-subject effects (main effect of estimation method, and interaction effects with estimation method) first. Furthermore, higher-order interactions are interpreted before main effects are interpreted. Some significant between-subject effects (significant main effects and interaction effects for other independent

variables) are interpreted subsequently. Again, interaction effects are presented and interpreted first before main effects. Finally, the compensatory and non-compensatory DIF indexes (CDIF and NCDIF) recorded are summarized across replications for the 14 items.

4.1.1 Within-Subject effects.

Table 4.1 presents the three-way interaction between sample size, latent mean differences and estimation method. Latent mean differences between the *reference* and *focal group* were manipulated such that the *focal group* came from a lower mean distribution for either one or both dimensions. This three-way interaction was significant for five out of the six dependent variables (except the non-DIF items), when sample size = 2000, but not for sample size = 1000.

When all 14 items were included in the test, the MGRM-DFIT method had a significantly higher likelihood of detecting DIF for all four latent mean difference conditions, when compared to the MG-CFA method. For all four latent mean difference conditions, the MGRM-DFIT method detected DIF for this dependent variable in all replications ($\hat{\pi} = 1.00$). However, the MG-CFA method detected DIF in only 19% of the replications ($\hat{\pi} = .19$) when there were no latent mean differences (i.e., no impact, Mean1 FG=0, Mean2 FG=-0), and when there was highest impact (Mean1 FG=-0.5, Mean2 FG=-0.5). When compared to these two conditions ($\hat{\pi} = .19$), the proportion of replications where DIF was detected increased for the MG-CFA method when Mean1 FG=0, Mean2 FG=-0.5 ($\hat{\pi} = .34$), and when Mean1 FG=0, Mean2 FG=-0.5 ($\hat{\pi} = .52$).

However, even in these conditions, the power of DIF detection for the MG-CFA method was significantly lower than the power of DIF detection for the MGRM-DFIT method.

For DIF item 37, the MG-CFA method ($\hat{\pi}=.24$) had a significantly higher likelihood in detecting DIF than the MGRM-DFIT method ($\hat{\pi}=.08$) when the *focal group* came from a lower ability distribution on the second dimension (Mean1 FG=0, Mean2 FG=-0.5). The MG-CFA method ($\hat{\pi}=.28$) also had a higher likelihood of detecting DIF than the MGRM-DFIT method ($\hat{\pi}=.12$) when the *focal group* came from a lower ability distribution on the first dimension (Mean1 FG=-0.5, Mean2 FG=0). However, the MG-CFA method and the MGRM-DFIT method did not differ in their likelihood of DIF detection for the other two latent mean difference conditions. The trend for the other three DIF items was similar to the trend just described.

In other words, for DIF item 38 also, the MG-CFA method ($\hat{\pi}=.25$) had a significantly higher likelihood in detecting DIF than the MGRM-DFIT method ($\hat{\pi}=.15$) when the *focal group* came from a lower ability distribution on the second dimension (Mean1 FG=0, Mean2 FG=-0.5). The MG-CFA method ($\hat{\pi}=.33$) also had a higher likelihood of detecting DIF than the MGRM-DFIT method ($\hat{\pi}=.22$) when the *focal group* came from a lower ability distribution on the first dimension (Mean1 FG=-0.5, Mean2 FG=0). Similarly, for item 39, the MG-CFA method ($\hat{\pi}=.26$) had a significantly higher likelihood in detecting DIF than the MGRM-DFIT method ($\hat{\pi}=.15$) when the *focal group* came from a lower ability distribution on the second dimension (Mean1 FG=0, Mean2 FG=-0.5). The MG-CFA method ($\hat{\pi}=.35$) also had a higher likelihood of detecting DIF than the MGRM-DFIT method ($\hat{\pi}=.23$) when the *focal group* came from a lower ability distribution on the first dimension (Mean1 FG=-0.5, Mean2 FG=0).

Table 4-1 Proportion of replications where DIF was detected for each Dependent Variable across the Sample size, Mean Difference, and Estimation method conditions.

Sample size	Mean Difference	Estimation Method	All 14 poly items	10 non-DIF poly items	Item 37	Item 38	Item 39	Item 40
1000	Mean1 FG = 0	MG-CFA	.12	.04	.05	.06	.06	.07
	Mean2 FG = 0	MGRM-DFIT	.59	.01	.06	.14	.14	.05
	Mean1 FG = 0	MG-CFA	.16	.11	.12	.12	.12	.13
	Mean2 FG = -0.5	MGRM-DFIT	.79	.02	.07	.14	.14	.05
	Mean1 FG = -0.5	MG-CFA	.28	.10	.13	.15	.16	.17
	Mean2 FG = 0	MGRM-DFIT	1.00	.02	.10	.19	.19	.10
	Mean1 FG = -0.5	MG-CFA	.10	.04	.05	.05	.06	.06
	Mean2 FG = -0.5	MGRM-DFIT	.79	.02	.10	.18	.18	.05
2000	Mean1 FG = 0	MG-CFA	.19*	.04	.06	.07	.08	.09
	Mean2 FG = 0	MGRM-DFIT	1.00**	.01	.08	.16	.16	.05
	Mean1 FG = 0	MG-CFA	.34*	.20	.24**	.25**	.26**	.26**
	Mean2 FG = -0.5	MGRM-DFIT	1.00**	.02	.08*	.15*	.15*	.06*
	Mean1 FG = -0.5	MG-CFA	.52*	.20	.28**	.33**	.35**	.37**
	Mean2 FG = 0	MGRM-DFIT	1.00**	.02	.12*	.22*	.23*	.15*
	Mean1 FG = -0.5	MG-CFA	.19*	.05	.07	.08	.09	.10
	Mean2 FG = -0.5	MGRM-DFIT	1.00**	.02	.11	.21	.21	.05

* *reference* condition(s) for each DV

** significant results (beta estimate and odds ratio described in Appendix A)

Finally, for item 40, the MG-CFA method ($\hat{\pi}=.26$) had a significantly higher likelihood in detecting DIF than the MGRM-DFIT method ($\hat{\pi}=.06$) when the *focal group* came from a lower ability distribution on the second dimension (Mean1 FG=0, Mean2 FG=-0.5). The MG-CFA method ($\hat{\pi}=.37$) also had a higher likelihood of detecting DIF than the MGRM-DFIT method ($\hat{\pi}=.15$) when the *focal group* came from a lower ability distribution on the first dimension (Mean1 FG=-0.5, Mean2 FG=0). However, the power of DIF detection did not differ between the two estimation methods for the other two latent mean difference conditions.

Since most of the significant results in Table 4.1 came from sample size = 2000, the results for this condition are graphically represented in Figure 4.1. It can be seen from Figure 4.1 that for the four DIF items, DIF detection increases for the MG-CFA method for the conditions where the *focal group* comes from a lower mean distribution in one of two dimensions (Mean1 FG=0, Mean2 FG=-0.5) and (Mean1 FG=-0.5, Mean2 FG=0). However, this figure clearly shows that DIF detection also increases for the non-DIF items in these two conditions. It can also be seen from Table 4.1 and Figure 4.1 that these two conditions are the cause for the overall inflated Type-I error rate for the MG-CFA method across all study conditions (since Type-I error is controlled in all other conditions for the MG-CFA method). Finally, for the overall test (all 14 items), DIF detection was higher across the board when N=2000. However, DIF detection was higher for the MGRM-DFIT method when compared to the MG-CFA method for this dependent variable.

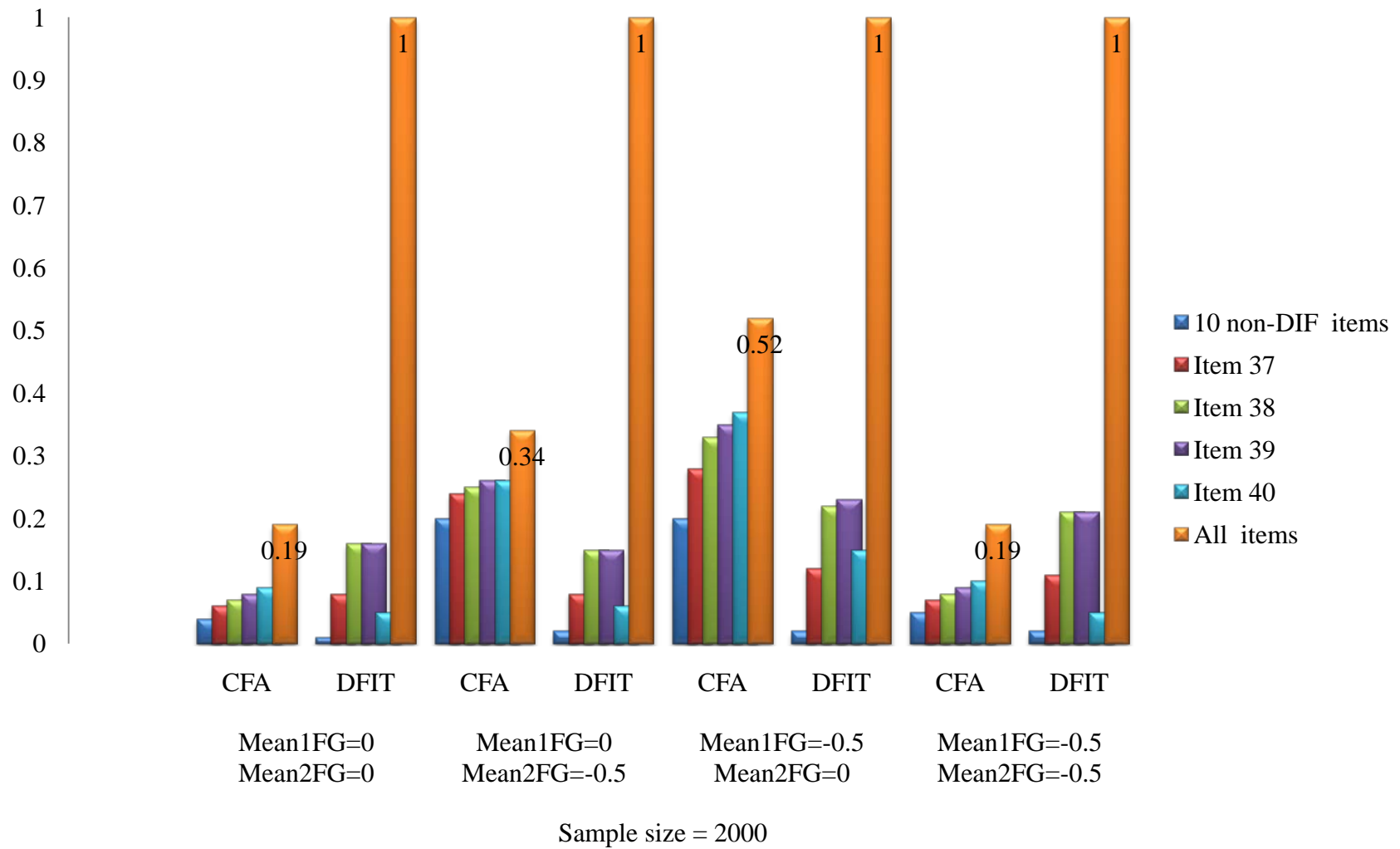


Figure 4-1 DIF detection for each DV across the Latent Mean Difference and Estimation method conditions when N=2000

Table 4.2 presents the two-way interaction between latent mean differences and estimation method. This effect was significant for the overall test (all 14 items), but not for the other dependent variables. In other words, there was no difference between the two estimation methods in detecting DIF for the non-DIF items, and for each DIF item individually. However, for the overall test, the MGRM-DFIT method ($\hat{\pi}=.89$) had a significantly higher likelihood of detecting DIF when compared to the MG-CFA method ($\hat{\pi}=.25$), when the *focal group* came from a lower mean distribution on the second dimension (Mean1 FG=0, Mean2 FG=-0.5). In addition, when the *focal group* came from a lower mean distribution on the first dimension (Mean1 FG=-0.5, Mean2 FG=0), the MGRM-DFIT method ($\hat{\pi}=1.00$) had a significantly higher likelihood of detecting DIF when compared to the MG-CFA method ($\hat{\pi}=.40$). Finally, when the *focal group* came from a lower mean distribution for both latent dimensions (Mean1 FG=-0.5, Mean2 FG=-0.5), the MGRM-DFIT method ($\hat{\pi}=.89$) had a significantly higher likelihood of detecting DIF compared to the MG-CFA method ($\hat{\pi}=.14$).

Table 4-2 Proportion of replications where DIF was detected for each Dependent Variable across the Mean Difference and Estimation Method conditions.

Mean Difference	Estimation Method	All 14 items	non-DIF items	Item 37	Item 38	Item 39	Item 40
Mean1 FG = 0	MG-CFA	.15	.04	.06	.06	.07	.08
Mean2 FG = 0	MGRM-DFIT	.79	.01	.07	.15	.15	.05
Mean1 FG = 0	MG-CFA	.25*	.16	.18	.19	.19	.20
Mean2 FG = -0.5	MGRM-DFIT	.89**	.02	.07	.15	.15	.06
Mean1 FG = -0.5	MG-CFA	.40*	.15	.21	.24	.26	.27
Mean2 FG = 0	MGRM-DFIT	1.00**	.02	.11	.21	.21	.12
Mean1 FG = -0.5	MG-CFA	.14*	.05	.06	.07	.08	.08
Mean2 FG = -0.5	MGRM-DFIT	.89**	.02	.10	.20	.20	.05

Table 4.3 presents the three-way interaction between uniform DIF, latent mean differences, and estimation method. This effect was significant for the overall test (with all 14 items). Uniform DIF was manipulated in two directions, either the items were designed to be more difficult for the *focal group* ($b_3=0$, $b_4=+0.5$; and $b_3=+0.5$, $b_4=+0.5$), or the items were manipulated to be more difficult for the *reference group* ($b_3=0$, $b_4=-0.5$; and $b_3=-0.5$, $b_4=-0.5$). In general, the MGRM-DFIT method had a higher likelihood of detecting DIF when compared to the MG-CFA method, when the items were manipulated to be more difficult for the *reference group* ($b_3=0$, $b_4=-0.5$; and $b_3=-0.5$, $b_4=-0.5$). Furthermore, for the two latent mean difference conditions, where the *focal group* came from a lower mean distribution on one of the two dimensions (Mean1 FG=0, Mean2 FG=-0.5 and Mean1 FG=-0.5, Mean2 FG=0), the MGRM-DFIT method ($\hat{\pi}$ ranging from .99 to 1.00) had a significantly higher likelihood of detecting DIF for the overall test when compared to MG-CFA method ($\hat{\pi}$ ranging from .24 to .58).

Specifically, for the conditions where it was more difficult for the *reference group* to score in the highest score category ($b_3=0$, $b_4=-0.5$):

- (i) The MGRM-DFIT method ($\hat{\pi}=1.00$) had a significantly higher likelihood of detecting DIF when the *focal group* came from a lower mean distribution on the second dimension (Mean1 FG=0, Mean2 FG=-0.5), when compared to the MG-CFA method ($\hat{\pi}=.24$).
- (ii) The MGRM-DFIT method ($\hat{\pi}=.99$) had a significantly higher likelihood of detecting DIF when the *focal group* came from a lower mean distribution on the first dimension (Mean1 FG=-0.5, Mean2 FG=0), when compared to the MG-CFA method ($\hat{\pi}=.46$).

Furthermore, for the conditions where it was more difficult for the *reference group* to score in the two highest score categories ($b_3=-0.5$, $b_4=-0.5$):

- (iii) The MGRM-DFIT method ($\hat{\pi}=1.00$) had a significantly higher likelihood of detecting DIF when the *focal group* came from a lower mean distribution on the second dimension (Mean1 FG=0, Mean2 FG=-0.5), when compared to the MG-CFA method ($\hat{\pi}=.27$).
- (iv) The MGRM-DFIT method ($\hat{\pi}=.99$) had a significantly higher likelihood of detecting DIF when the *focal group* came from a lower mean distribution on the first dimension (Mean1 FG=-0.5, Mean2 FG=0), when compared to the MG-CFA method ($\hat{\pi}=.58$).

In general, there was no difference between the two estimation methods in detecting DIF for the remaining dependent variables. Furthermore, it can be seen from Table 4.3 that the overall DIF detection rate increases for the MG-CFA method when the *focal group* comes from a lower mean distribution on either one of (but not both) the two latent dimensions. The DIF detection rate increases for the MG-CFA method even for the non-DIF items, thereby resulting in the overall inflation of Type-I error rate for this estimation method. However, this trend does not continue for the CFA method when the *focal group* comes from a lower ability distribution on both latent dimensions.

Table 4-3 Proportion of replications where DIF was detected for each Dependent Variable across the Uniform DIF, Latent Mean Difference, and Estimation Method conditions.

Uniform DIF	Mean Difference	Estimation Method	All items	non-DIF items	Item 37	Item 38	Item 39	Item 40
b3=0 b4=+0.5	Mean1 FG = 0	MG-CFA	.08	.04	.04	.04	.05	.05
	Mean2 FG = 0	MGRM-DFIT	.59	.01	.05	.14	.14	.04
	Mean1 FG = 0	MG-CFA	.23	.15	.17	.18	.18	.19
	Mean2 FG = -0.5	MGRM-DFIT	.80	.02	.05	.14	.14	.08
	Mean1 FG = -0.5	MG-CFA	.27	.15	.18	.19	.20	.19
	Mean2 FG = 0	MGRM-DFIT	.99	.02	.09	.19	.20	.08
	Mean1 FG = -0.5	MG-CFA	.07	.04	.05	.05	.06	.06
	Mean2 FG = -0.5	MGRM-DFIT	.79	.01	.09	.17	.17	.06
b3=+0.5 b4=+0.5	Mean1 FG = 0	MG-CFA	.09	.05	.06	.06	.06	.07
	Mean2 FG = 0	MGRM-DFIT	.58	.01	.06	.14	.15	.04
	Mean1 FG = 0	MG-CFA	.26	.16	.19	.19	.20	.20
	Mean2 FG = -0.5	MGRM-DFIT	.78	.01	.05	.13	.13	.04
	Mean1 FG = -0.5	MG-CFA	.29	.15	.18	.20	.19	.22
	Mean2 FG = 0	MGRM-DFIT	.99	.02	.09	.19	.19	.04
	Mean1 FG = -0.5	MG-CFA	.10	.05	.06	.06	.07	.07
	Mean2 FG = -0.5	MGRM-DFIT	.81	.02	.09	.17	.18	.04
b3=0 b4=-0.5	Mean1 FG = 0	MG-CFA	.16	.04	.06	.07	.07	.08
	Mean2 FG = 0	MGRM-DFIT	1.00	.01	.08	.16	.16	.06
	Mean1 FG = 0	MG-CFA	.24*	.16	.18	.19	.19	.20
	Mean2 FG = -0.5	MGRM-DFIT	1.00**	.02	.09	.16	.16	.06
	Mean1 FG = -0.5	MG-CFA	.46*	.15	.21	.25	.28	.29
	Mean2 FG = 0	MGRM-DFIT	.99**	.02	.12	.22	.22	.06
	Mean1 FG = -0.5	MG-CFA	.14	.05	.06	.06	.06	.08
	Mean2 FG = -0.5	MGRM-DFIT	.98	.02	.12	.21	.22	.05
b3=-0.5 b4=-0.5	Mean1 FG = 0	MG-CFA	.28	.04	.07	.09	.11	.12
	Mean2 FG = 0	MGRM-DFIT	1.00	.02	.09	.17	.16	.06
	Mean1 FG = 0	MG-CFA	.27*	.15	.18	.18	.19	.20
	Mean2 FG = -0.5	MGRM-DFIT	1.00**	.02	.10	.16	.15	.07
	Mean1 FG = -0.5	MG-CFA	.58*	.15	.25	.32	.36	.39
	Mean2 FG = 0	MGRM-DFIT	.99**	.03	.13	.24	.23	.06
	Mean1 FG = -0.5	MG-CFA	.25	.05	.08	.10	.12	.12
	Mean2 FG = -0.5	MGRM-DFIT	.98	.03	.13	.22	.22	.06

Table 4.4 presents the two-way interaction between uniform DIF and estimation method. This effect was significant for the overall test (all 14 items). When it was more difficult for the *reference group* to score a 5, the MGRM-DFIT method ($\hat{\pi}=1.00$) had a significantly higher likelihood of detecting DIF compared to the MG-CFA method ($\hat{\pi}=.25$). In addition, the MGRM-DFIT method ($\hat{\pi}=1.00$) had a significantly higher likelihood of detecting DIF when it was more difficult for the *reference group* to score a 4 and a 5, when compared to the MG-CFA method ($\hat{\pi}=.35$).

Table 4-4 Proportion of replications where DIF was detected for each Dependent Variable across the Uniform DIF and Estimation method conditions.

Uniform DIF	Estimation Method	All 14 items	non-DIF items	Item 37	Item 38	Item 39	Item 40
b3 = 0	MG-CFA	.16	.10	.11	.12	.12	.13
b4 = +0.5	MGRM-DFIT	.79	.01	.07	.16	.16	.06
b3 = +0.5	MG-CFA	.18	.10	.12	.13	.13	.14
b4 = +0.5	MGRM-DFIT	.79	.01	.07	.16	.16	.06
b3 = 0	MG-CFA	.25*	.10	.13	.14	.15	.16
b4 = -0.5	MGRM-DFIT	1.00**	.02	.10	.19	.19	.08
b3 = -0.5	MG-CFA	.35*	.10	.15	.17	.19	.21
b4 = -0.5	MGRM-DFIT	1.00**	.02	.11	.20	.19	.08

The results from Table 4.4 are graphically represented in Figure 4.2. It can be seen from Figure 4.2 that for five of the dependent variables (non-DIF items and each of the 4 DIF items) DIF detection was low, and did not differ significantly across study conditions. However, for the entire test (with all 14 items), DIF detection was significantly higher across the board. Furthermore, for this dependent variable, DIF

detection was significantly higher for the MGRM-DFIT method when compared to the MG-CFA method across the four uniform DIF conditions.

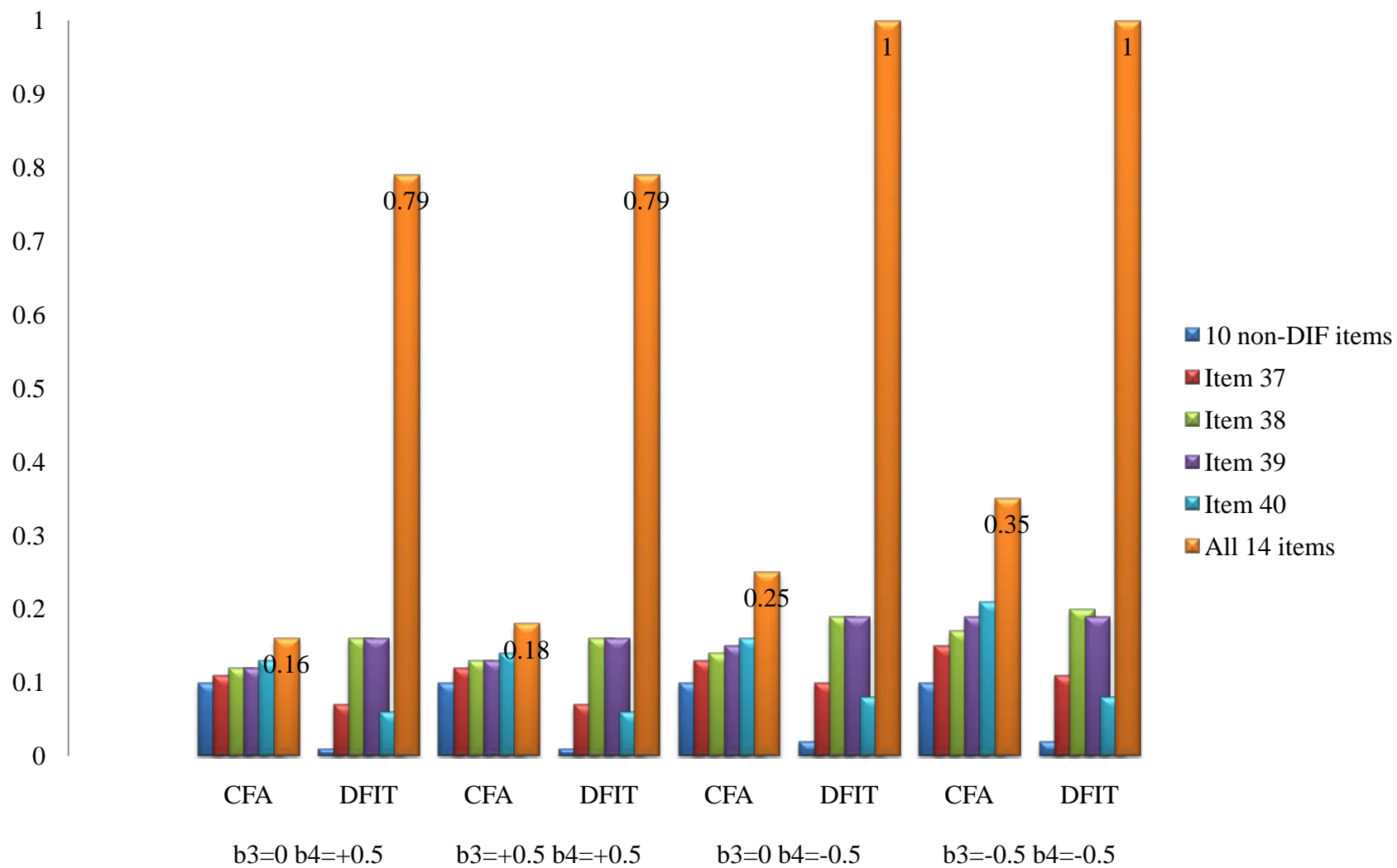


Figure 4-2 DIF detection for each Dependent Variable across the Uniform DIF and Estimation method conditions

Overall, from Tables 4.3 and 4.4, it can be deduced that DIF detection was higher when the items were manipulated to be more difficult for the *reference group*. The *focal group* came from a lower ability distribution (for three of four latent mean difference conditions). Though the effect size was not significant, it can be seen from Table 4.3 that DIF detection is significantly high for the MGRM-DFIT method when the *focal group* comes from a lower ability distribution on both latent dimensions. For all these conditions, *impact* (true distributional differences) is in the opposite direction of manipulated DIF. It is therefore likely that DIF detection was exaggerated in these conditions where the item is manipulated to be harder for the *reference group*.

Table 4-5 Proportion of replications where DIF was detected for each Dependent Variable in the control condition across the two Estimation methods.

Estimation Method	All 14 items	non-DIF items	Item 37	Item 38	Item 39	Item 40
MG-CFA	.02	.02	.02	.02	.02	.02
MGRM-DFIT	.15	.02	.07	.09	.13	.05

Table 4-6 Proportion of replications where DIF was detected for each Dependent Variable averaged across the study conditions for the two Estimation methods.

Estimation Method	All 14 items	non-DIF items	Item 37	Item 38	Item 39	Item 40
MG-CFA	.24*	.10	.13	.14	.15	.16*
MGRM-DFIT	.89**	.02	.09	.18	.18	.07**

Overall Type-I error rate and power for the two estimation methods are presented in Tables 4.5 and 4.6. One hundred replications of a control condition were simulated,

where sample size was fixed at 2000, sample size ratio was fixed at 50/50, and all other independent variables were held constant at zero. The baseline null hypothesis rejection rates were established for the two estimation methods in this condition. Table 4.5 presents the proportion of replications, for each estimation method, where DIF was detected for each of the dependent variables in the control condition. Rejecting the null hypothesis in these conditions (where no DIF was simulated) is also an indicator of Type-I error rate. To this effect, the Type-I error rate is slightly inflated for the MGRM-DFIT method in the control condition.

Table 4.6 presents the proportion of replications where DIF was detected for each estimation method averaged across the study conditions. The main effect of estimation method was significant for two of the six dependent variables. The MGRM-DFIT method ($\hat{\pi}=.89$) had a higher likelihood of detecting DIF in the total test (all 14 items) than the MG-CFA method ($\hat{\pi}=.24$). However, the MGRM-DFIT method ($\hat{\pi}=.07$) had a lower likelihood of detecting DIF for item 40 (DIF item) than the MG-CFA method ($\hat{\pi}=.16$). It is apparent from Table 4.6, however, that the power of DIF detection for the overall test and the individual DIF items is considerably different for the MGRM-DFIT method. Though the DFIT method seems to perform as a good multivariate test, it is not as efficient as a univariate test.

4.1.2 Between-Subject effects.

The two-way interaction between sample size and latent mean differences was significant for all four DIF items. These results are presented in Table 4.7. In general, when the *focal*

group came from a lower ability distribution on either one of the latent dimensions, there was a significantly higher likelihood of detecting DIF for the N=2000 condition compared to the N = 1000 condition. It should be recalled that the power of DIF detection increased for these two latent mean difference conditions for the MG-CFA method, and not for the MGRM-DFIT method. The results from the logistic regression with beta parameter estimates and odds ratio are presented in Appendix A for this table.

Table 4-7 Proportion of replications where DIF was detected for each Dependent Variable across the Mean Difference and Sample size conditions.

Mean Difference	Sample size	All 14 items	non-DIF items	Item 37	Item 38	Item 39	Item 40
Mean1 FG = 0	1000	.36	.03	.06	.10	.10	.06
Mean2 FG = 0	2000	.59	.03	.07	.12	.12	.07
Mean1 FG = 0	1000	.48	.06	.09*	.13*	.13*	.09*
Mean2 FG = -0.5	2000	.67	.11	.16**	.20**	.20**	.16**
Mean1 FG = -0.5	1000	.64	.06	.12*	.17*	.18*	.13*
Mean2 FG = 0	2000	.76	.11	.20**	.28**	.29**	.26**
Mean1 FG = -0.5	1000	.44	.03	.08	.12	.12	.06
Mean2 FG = -0.5	2000	.59	.04	.09	.14	.15	.08

The Main effects of Uniform DIF and Latent mean differences were significant for the overall test (all 14 items). The mean DIF detection rate for each of the six dependent variables across the Latent mean difference and Uniform DIF conditions are presented in Tables 4.8 and 4.9, respectively. Table 4.8 shows that, there was a significantly higher power of DIF detection ($\hat{\pi}=.70$) when the *focal group* came from a lower mean distribution on the first dimension (Mean1 FG=0, Mean2 FG=-0.5) when

compared to the condition ($\hat{\pi}=.48$) with equal latent means across groups (Mean1 FG=0, Mean2 FG=0). Table 4.9 shows that, there was a significantly higher power of DIF detection ($\hat{\pi}=.67$) when the items were more difficult for the *reference group* ($b3=-0.5$, $b4=-.5$), when compared to the condition ($\hat{\pi}=.47$) where it was more difficult for the *focal group* to score a 5 ($b3=0$, $b4=+0.5$).

Table 4-8 Proportion of replications where DIF was detected for each Dependent Variable across the Mean Difference conditions.

Mean Difference	All items	14 non-DIF items	Item 11	Item 12	Item 13	Item 14
Mean1 FG = 0	.48*	.03	.06	.11	.11	.07
Mean2 FG = 0						
Mean1 FG = 0	.57	.09	.13	.17	.17	.13
Mean2 FG = -0.5						
Mean1 FG = -0.5	.70**	.09	.16	.22	.23	.19
Mean2 FG = 0						
Mean1 FG = -0.5	.51	.03	.08	.13	.14	.07
Mean2 FG = -0.5						

Table 4-9 Proportion of replications where DIF was detected for each Dependent Variable across the Uniform DIF conditions.

Uniform DIF	All 14 items	non-DIF items	Item 11	Item 12	Item 13	Item 14
b3 = 0	.47*	.06	.09	.14	.14	.09
b4 = +0.5						
b3 = +0.5	.57	.06	.10	.14	.15	.10
b4 = +0.5						
b3 = 0	.62	.06	.11	.17	.17	.12
b4 = -0.5						
b3 = -0.5	.67**	.06	.13	.18	.19	.14
b4 = -0.5						

4.1.3 CDIF and NCDIF.

Finally, the CDIF and NCDIF values for each of the 14 polytomous items were computed for each replication in each condition. The CDIF and NCDIF values for all items averaged across replications are presented in Table 4.10. It should be recalled from Eq. (29) that the CDIF measure is computed as a covariance between the d_i (d_i =difference in expected item scores between RG and FG) for the given item and the total test. On the other hand, NCDIF values are calculated as the sum of the mean and variance of d_i for each item in isolation. Therefore, larger CDIF and NCDIF values reflect a higher amount of DIF in each item.

Table 4-10 CDIF and NCDIF values for the 14 items across study conditions

Item #	CDIF	NCDIF
Item 27	33.90	3.86
Item 28	38.24	3.64
Item 29	35.40	3.65
Item 30	33.14	3.64
Item 31	31.16	3.61
Item 32	29.40	3.59
Item 33	35.57	3.38
Item 34	34.88	3.11
Item 35	34.57	3.51
Item 36	32.87	3.49
Item 37	20.87	4.45
Item 38	19.27	5.41
Item 39	17.98	5.37
Item 40	25.48	4.04

It can be seen from Table 4.10 that overall NCDIF values are larger for the 4 DIF items, when compared to the rest of the items in the test. However, CDIF values are actually smaller for the 4 DIF items when compared to the rest of the items in the test. This means that across the replications and the study conditions, CDIF was not as consistent as NCDIF, and there might have been several outlier observations for both DIF and non-DIF items.

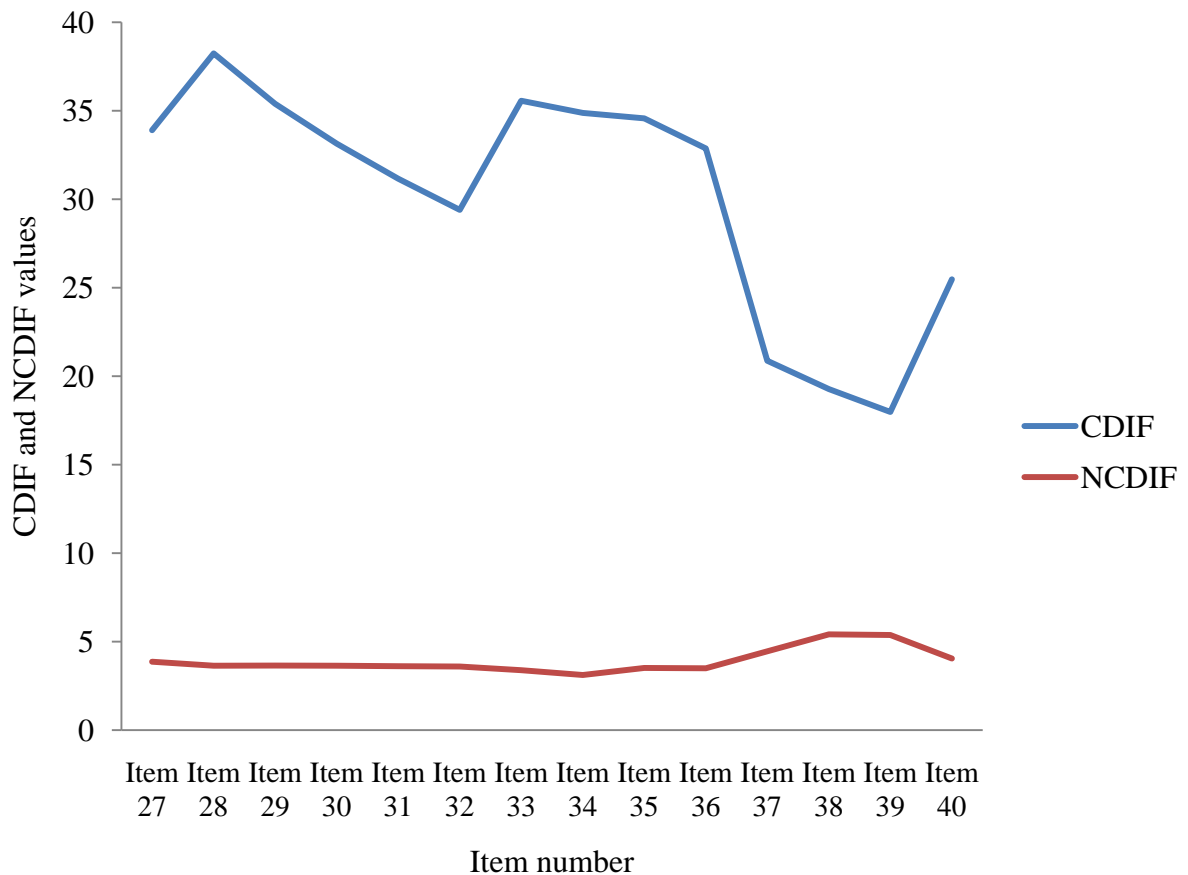


Figure 4-3 CDIF and NCDIF values for all 14 items averaged across replications

In addition, overall DIF detection was somewhat lower for items 37 and 40 (smaller NCDIF values) than for items 38 and 39 (larger NCDIF values). This reiterates

what was found from the power of DIF detection for these four items presented across various independent variables in Tables 4.1 through 4.9. The CDIF and NCDIF values are also presented graphically in Figure 4.3. Again, it can be seen from this figure that, compared to the rest of the items, NCDIF values are larger for the 4 DIF items. However, the NCDIF values are slightly lower for items 37 and 40 when compared to items 38 and 39.

5.0 DISCUSSION

This study was intended to evaluate the robustness of DIF detection for multidimensional polytomous items using two different estimation methods, MG-CFA and MGRM-DFIT. A simulation study across 960 study conditions was performed. The purpose of this study was to establish the Type-I error rate and Power of DIF detection for the MG-CFA and MGRM-DFIT estimation methods across the study conditions. This study also aimed to determine the pattern of differences in the Type-I error rate and Power of the two estimation methods among the levels of the independent variables considered in this study(*sample size, sample size ratio, type of DIF, DIF direction, and differences in latent distribution*).

This chapter summarizes some of the major findings from this study. In addition, an overview of some of the limitations of the study overall, and each of the estimation methods, in particular are discussed. Finally, some practical implications of the findings and directions for future research are provided.

5.1 SUMMARY OF MAJOR FINDINGS

5.1.1 Type-I error rates and empirical power for the MGRM-DFIT estimation method summarized across study conditions.

The MGRM-DFIT method was found to consistently control Type-I error rate under alpha across all study conditions. For almost all study conditions, the rate of DIF detection in the non-DIF items for this method was less than 2%. Though the MGRM-DFIT method demonstrated high power in detecting DIF for the overall test, it had lower power in detecting DIF for each DIF item individually. This can be partly explained by the compensatory nature of the DFIT method. The DFIT method starts from detecting DIF in the overall test (DTF), and individual items are then compared for their covariance with the overall test. The overall DTF (Differential Test Functioning) in the DFIT method is computed by taking the amount of DIF in each item into account (see Eq. 25). When the chi-square test for each DIF item is performed individually, all other DIF items were already removed from the test.

Furthermore, the DFIT method uses sample size (of the *reference* or *focal group*) as the degrees of freedom for the chi-square test. Therefore, the degrees of freedom (*df*) for all the tests are the same for the MGRM-DFIT method, as opposed to the MG-CFA method where the degrees of freedom corresponds to the number of parameters used in the respective test. Additionally, it can be seen from Table 4.5 that the DFIT method had a higher rejection rate (15%) for the overall test even for a control condition when no DIF was introduced. Therefore, the chisquare test could be extremely sensitive for the overall

test, due to the degrees of freedom (*df*) used. These above reasons might partly explain why there is higher power of DIF detection for the overall test (when all 4 DIF items are present) than when only one DIF item is present.

Since the DIF detection for the overall test and for the individual items was very different for the DFIT method, this method did not consistently detect DIF for all dependent variables. However, as evident from the results, though the MG-CFA method had relatively lower power in detecting DIF for the overall test, the DIF detection for the test, and individual items were consistent for this method. Therefore, it looks like the MGRM-DFIT method works well as a multivariate test, but not as a univariate test, while the MG-CFA method performs consistently in the univariate and multivariate scenarios.

Finally, the MGRM-DFIT method had higher power of DIF detection when the items were manipulated to be more difficult for the *reference group* (see Tables 4.3 and 4.4). The *focal group* came from a lower ability distribution for three of four latent mean difference conditions. For these conditions, *impact* (true distributional differences) is in the opposite direction of manipulated DIF. Previous studies (Bolt, 2002; Flowers, Oshima & Raju, 1999; Stark, Chernyshenko & Drasgow, 2006) have found that impact results in higher number of false positives. It is therefore likely that DIF detection was exaggerated in these conditions, especially since the items were manipulated to be harder for the *reference group*, but impact was manipulated such that the *focal group* comes from a lower mean distribution.

5.1.1.1 CDIF and NCDIF.

In general, higher CDIF and NCDIF values are indicative of the presence of DIF. Overall, compared to the non-DIF items, NCDIF values are larger, and CDIF values are smaller (see Table 4.10 and Fig. 4.3) for the 4 DIF items. This means that across the replications and the study conditions, CDIF was not as consistent as NCDIF, and there might have been several outlier observations for both DIF and non-DIF items. Previous research (Fleer, 1993; Flowers, Oshima & Raju, 1999; Oshima, Raju & Flowers, 1997) has found that the CDIF was not as stable as the NCDIF. They found that CDIF erroneously identified 48% of the non-DIF items as DIF and only identified 50% of the DIF items correctly (Flowers, Oshima & Raju, 1999). The authors attribute the erratic detection rate for CDIF to possible estimation and linking errors accumulated across the test (Flowers, Oshima & Raju, 1999). This is a possible limitation of the method, and since DTF is calculated as the sum of CDIF values, this could also possibly explain some of the spuriously high DIF detection rates for the overall test for the MGRM-DFIT method.

5.1.2 Type-I error rates and empirical power for the MG-CFA estimation method summarized across study conditions.

The MG-CFA method demonstrated a slightly inflated Type-I error rate. In other words, it had a higher (than α) probability of detecting DIF for the non-DIF items. However, as will be discussed subsequently, it looks like a couple of study conditions (with inflated

Type-I error rate) contributed to this overall trend. Otherwise, the MG-CFA method was found to control Type-I error rate under alpha for other study conditions.

However, the MG-CFA method demonstrated lower power across all study conditions. With the exception of a few study conditions, DIF detection rate was less than 25% for the MG-CFA method across most study conditions. This could partly be explained by the low magnitude of DIF that was manipulated in the ' α/λ ' parameter. Previous research (Flowers, Oshima & Raju, 1999; Meade, Lautenschlager & Johnson, 2006; Stark, Chernyshenko & Drasgow, 2006), has repeatedly found that items with larger DIF magnitudes are more easily detected. However, due to the computational limitations of Eq. (9) and (10) (see Chapter III, Study Design), the DIF magnitude for the ' α ' parameter was limited to 0.2 in this study. The CFA method is more sensitive to changes in the ' λ ' parameter, and this small magnitude of DIF might have resulted in the lower empirical power of DIF detection for the MG-CFA method (see further elaborations under study limitations).

DIF detection for the MG-CFA method seemed to increase as sample size and impact increased (see Table 4.1 and Fig. 4.1). That is, when the *focal group* came from a lower mean distribution in one of two dimensions, and sample size = 2000, DIF detection increased for the MG-CFA method. The result with regard to sample size is in contrast to previous findings (Gonzalez-Roma, Hernandez & Gomez-Benito, 2006; Stark, Chernyshenko & Drasgow, 2006), where the MG-CFA method was found to produce stable and consistent results at smaller sample sizes. However, it should be noted that these previous studies investigated unidimensional models. It is likely that larger sample sizes (larger than those used in the current study) are required to produce stable results

for multidimensional polytomous models (since the total number of parameters estimated per item was large).

Furthermore, for the four DIF items, DIF detection was found to increase for the MG-CFA method in conditions where the *focal group* came from a lower mean distribution in one of two dimensions (Mean1 FG=0, Mean2 FG=-0.5) and (Mean1 FG=-0.5, Mean2 FG=0). However, Fig. 4.1 clearly shows that DIF detection also increases for the non-DIF items in these two conditions. It can be seen from Table 4.1 and 4.2 that these two conditions are the cause for the overall inflated Type-I error rate for the MG-CFA method. Unequal latent distributions between the *reference* and *focal groups* have been found to inflate Type-I error rate in general (Gonzalez-Roma, Hernandez & Gomez-Benito, 2006; Wu & Lei, 2009).

We should, therefore, have our reservations in assuming that the CFA method demonstrated higher power in detecting DIF in these conditions. It is likely that the presence of impact in one dimension possibly caused the MG-CFA model to detect DIF at a higher rate. Previous studies (Bolt, 2002; Flowers, Oshima & Raju, 1999; Stark, Chernyshenko & Drasgow, 2006) have found that impact results in higher number of false positives. However, it is interesting to note that this trend (of higher DIF detection) did not continue for the condition where impact was at the highest, i.e., when the *focal group* came from a lower ability distribution for both dimensions (Mean1 FG=-0.5, Mean2 FG=-0.5).

5.1.3 Effect of Independent Variables.

Overall, sample-size, latent mean differences, and uniform DIF were found to influence the rate of DIF detection across study conditions for the two estimation methods. However, sample-size ratio (SSR) and non-uniform DIF did not have an impact on DIF detection across study conditions. In general, disparate sample sizes should not make a difference for the CFA method. However, previous research (Gonzalez-Roma, Hernandez & Gomez-Benito, 2006) has found that unequal sample sizes results in larger Type-I error rates for the IRT-based methods. It is interesting to note that there was no significant difference in DIF detection between equal SSR of 50/50 and unequal SSR of 80/20 (which would be considered highly disparate). The low magnitudes of DIF simulated overall resulted in lower power in DIF detection across most study conditions. Therefore, it is likely that the differences between SSR of 50/50 and 80/20 were masked due to the overall low power in DIF detection.

Meade and Lautenschlager (2004a) claimed that differences in the ‘a’ parameter (or *non-uniform* DIF) would be picked up equally well under both the IRT and CFA methods. However, differences due to *non-uniform* DIF were not picked up in the current study. Furthermore, researchers (Gonzalez-Roma, Hernandez & Gomez-Benito, 2006; Stark, Chernyshenko & Drasgow, 2006) have found that power for detecting *non-uniform* DIF was significantly related to sample size. In addition, DIF detection was also influenced by DIF magnitude (Flowers, Oshima & Raju, 1999; Meade, Lautenschlager & Johnson, 2006; Stark, Chernyshenko & Drasgow, 2006).

Therefore, in order for *non-uniform* DIF (differences in the ‘a’ parameter) to be detected well, larger magnitude of DIF should have been simulated in the ‘a’ parameter. However, as has been pointed out previously (and will be elaborated subsequently), DIF magnitude for the ‘a’ parameter had to be limited to 0.2 due to computational limitations in this study. Furthermore, this study did not have conditions where only *non-uniform* DIF was manipulated (with no *uniform* DIF). Crossing *non-uniform* DIF with a condition where no DIF is introduced in the ‘b’ parameter, or incorporating an incomplete crossed design might have helped pick up differences due to *non-uniform* DIF better.

5.2 LIMITATIONS, FUTURE DIRECTIONS AND PRACTICAL IMPLICATIONS

A Monte Carlo study was used to address the proposed research questions in this study. Multidimensional polytomous items are still not predominantly used within state assessments. Therefore, there is a dearth in the availability of real data sets for multidimensional polytomous types of items, and the current estimation methods could not be applied to real state assessments. There are a number of extraneous factors that could potentially affect the behavior of real datasets, and these factors are not replicated well in a Monte Carlo simulation. This largely limits the generalizability of the results from the estimation methods, since it has not yet been applied to real data. Researchers should focus on developing tests and assessments with more multidimensional polytomous types of items, so that the practicability of these estimation methods could be determined.

Furthermore, though the study conditions were carefully designed to include a variety of conditions encountered in real performance data, the results may not generalize to other situations not considered in the current study. In general, there were some limitations to some of the factors held constant in this study. First, the number of response categories was limited to five in this study (only a 5-point scale was used), and only a graded model was used within the current study. However, state assessments could use a variety of polytomous scale-points in their rubrics, and could be suitable for other types of polytomous models (such as the partial-credit model). For example, the NAEP uses a 3-point grading system for its polytomous assessments, and these response categories were not modeled in this study. Therefore, future research should focus on assessing the robustness of these estimation methods for other types of polytomous scales, and for assessments that do not fit the graded model.

Second, the magnitude of DIF for the ' α ' parameter was limited to 0.2. The Mplus baseline estimates were used in order to compute the parameters for the MGRM-DFIT procedure, and the DIF magnitude for the ' λ ' parameter had to be limited in order for Eq. (9) and (10) to be estimable (see Chapter III, Study Design). Therefore, due to the computational limitations of the MGRM parameters (Kannan & Kim, 2009), and the already high ' α_1 ' and ' ρ ' parameter values used in the study, magnitude of DIF for the ' α ' parameter had to be limited to 0.2. Furthermore, since values for the ' α_2 ' were generated from a $N(0,0.1)$ distribution, the secondary factor loading were considerably small (of $< \sim .15$). This does not reflect a truly multidimensional model. CFA is more sensitive to changes in the ' λ ' parameter, and metric invariance is first established within the CFA method. Introducing a DIF of 0.2 in the ' α ' parameter translates to a DIF of

~0.12 in the ' λ ' parameter, and this small magnitude of DIF might have resulted in the lower empirical power in DIF detection for the MG-CFA method.

Future research, should assess larger DIF magnitudes in the ' α/λ ' parameter, and larger magnitudes of ' α_2 ' values, especially when only the MG-CFA method is being used. The computational limitation mentioned above does not hold when only the MG-CFA model is estimated. The limitation was inherent to the nature of this study, since parameters estimated using the CFA baseline model were transformed and used for the MGRM-DFIT method. A follow-up simulation should be performed with only the MG-CFA estimation method, wherein larger DIF magnitudes for the ' α/λ ' parameter are simulated, and DIF magnitudes are varied for both the ' α/λ ' and the ' β/τ ' parameters.

In addition to design limitations, there were some additional limitations to the MGRM-DFIT estimation method and the Mplus software which made comparisons between the two estimation methods limiting in some ways. First, the DFIT procedure does not provide parameter-level information. DIF is detected at the test level, and some item-level information (CDIF and NCDIF) are available for practical usage. However, there is no way to test if a given ' α ' or ' β ' parameter within an item was detected with DIF. On the other hand, the MG-CFA method provides modification indices for each parameter within the item. These modification indices are chi-square distributed, and can be tested for significance independently. Due to the lack of a comparable test within the DFIT method, these modification indices were not used in the current study.

Second, the CDIF values generated for each item by the MGRM-DFIT procedure do not have an independent significance test. In a practical sense, DTF would be tested for the entire test, and items with the largest CDIF values would be removed one at a

time, until the chi-square test for DTF (with the remaining items) is not significant. However, from a simulation perspective, each item could not be manually compared and removed for each replication. Therefore, model comparisons, with different number of items in nested models, were used for the DFIT method. This forced us to use model comparisons with various constrained models for the MG-CFA procedure. Unfortunately, the Mplus software takes around one to two minutes for each model to converge. Therefore, with 7 models (one baseline and 6 constrained models) for each replication, the amount of time taken per replication, per condition was fairly high.

From a practical stand-point, multiple model comparisons is not as much of an obstacle for either estimation method. For the MGRM-DFIT method, CDIF values will be estimated for each item, and items with larger CDIF values can be removed one at a time to test for DTF. Additionally, with applied data (a single data set), the amount of time Mplus takes for multiple model comparisons is not significant. Even if 20 model comparisons were to be performed, it would only take a total of anywhere between 20-40 minutes. Furthermore, when the MG-CFA method is used with applied data, model modification indices for each parameter can be examined. The modification indices are chi-square distributed, and can be used to determine the amount of DIF present in each parameter. Therefore, item- and parameter-level information can be obtained from the MG-CFA method.

Finally, the MGRM-DFIT procedure has not been implemented as part of any software package. This further limits the practical usability of this estimation method. Users would have to either implement the SAS macros created in this study, or write several lines of codes themselves in order to assess DIF for Multidimensional polytomous

types of items using the DFIT procedure. Furthermore, parameter estimation for the MGRM is still not widely available as part of commonly used software packages. This is one of the reasons why the Mplus software (the MG-CFA method's baseline model) had to be used to estimate item parameters and ability estimates required by the MGRM-DFIT procedure. The authors (Raju, et al., 1995; 1997; 1999) who have diligently worked on promoting the DFIT method by developing the DFIT software, should try and incorporate the MGRM-DFIT method to their existing software. In addition, other researchers might want to incorporate the MGRM-DFIT into a practically usable software that also estimates MGRM item parameters and ability estimates.

In general, the MG-CFA method is recommended for DIF detection with multidimensional polytomous types of items. The MG-CFA method performs more consistently in detecting DIF, both as a univariate test and as a multivariate test. However, the MGRM-DFIT method has high power as a multivariate test, but fails to detect DIF in individual items. Furthermore, the MG-CFA method is easily available as part of several commonly used software packages, while the MGRM-DFIT method is not.

APPENDIX A

RESULTS FROM THE LOGISTIC REGRESSION

The proportion of replications where DIF was detected by each estimation method for each Dependent variable are presented in Table 4.1 through 4.9. Results from the logistic regression with beta parameter estimates and odds ratio for the statistically significant effects are presented in this Appendix.

A.1 WITHIN-SUBJECT EFFECTS

Table 4.1 presents the three-way interaction between sample size, latent mean differences and estimation method. This effect was significant for five out of the six dependent variables (except the non-DIF items). For the dependent variable with all 14 items, the MGRM-DFIT method had a significantly higher likelihood of detecting DIF for all four latent mean difference conditions when sample size=2000. Specifically:

- (i) When the *focal group* came from a comparable mean distribution (i.e., no impact; Mean1 FG=0, Mean2 FG=0), the MGRM-DFIT method ($\hat{\pi}=1.00$) had a significantly higher likelihood ($B=.172$, $OR=1.188$, $p < .001$) of detecting DIF when compared to the MG-CFA method ($\hat{\pi}=.19$).
- (ii) When the *focal group* came from a lower mean distribution on the second dimension (Mean1 FG=0, Mean2 FG=-0.5), the MGRM-DFIT method ($\hat{\pi}=1.00$)

- had a significantly higher likelihood ($B=.110$, $OR=1.116$, $p < .001$) of detecting DIF when compared to the MG-CFA method ($\hat{\pi}=.34$).
- (iii) When the *focal group* came from a lower mean distribution on the first dimension (Mean1 FG=-0.5, Mean2 FG=0), the MGRM-DFIT method ($\hat{\pi}=1.00$) had a significantly higher likelihood ($B=.164$, $OR=1.178$, $p < .001$) of detecting DIF when compared to the MG-CFA method ($\hat{\pi}=.52$).
 - (iv) Finally, when the *focal group* came from a lower mean distribution on both dimensions (Mean1 FG=-0.5, Mean2 FG=-0.5), the MGRM-DFIT method ($\hat{\pi}=1.00$) had a significantly higher likelihood ($B=.172$, $OR=1.188$, $p < .001$) of detecting DIF when compared to the MG-CFA method ($\hat{\pi}=.19$).

For DIF items 37 through 40, when sample size=2000, the MG-CFA method had a significantly higher likelihood than the MGRM-DFIT method in detecting DIF when the *focal group* came from a lower mean distribution on either of the two latent dimension second dimension, i.e., (Mean1 FG=0, Mean2 FG=-0.5) and (Mean1 FG=-0.5, Mean2 FG=0).

For item 37, the MG-CFA method ($\hat{\pi}=.24$) had a significantly higher likelihood ($B=.114$, $OR=1.121$, $p < .001$) of detecting DIF when the *focal group* came from a lower mean distribution on the second dimension (Mean1 FG=0, Mean2 FG=-0.5), compared to the *reference* MGRM-DFIT method ($\hat{\pi}=.08$). In addition, the MG-CFA method ($\hat{\pi}=.28$) had a significantly higher likelihood ($B=.126$, $OR=1.134$, $p < .001$) of detecting DIF when the *focal group* came from a lower mean distribution on the first dimension (Mean1 FG=-0.5, Mean2 FG=0), compared to the MGRM-DFIT method ($\hat{\pi}=.12$).

For item 38, the MG-CFA method ($\hat{\pi}=.25$) had a significantly higher likelihood ($B=.119$, $OR=1.126$, $p < .001$) of detecting DIF when the *focal group* came from a lower mean distribution on the second dimension (Mean1 FG=0, Mean2 FG=-0.5), when compared to the MGRM-DFIT method ($\hat{\pi}=.15$). In addition, the MG-CFA method ($\hat{\pi}=.33$) had a significantly higher likelihood ($B=.149$, $OR=1.161$, $p < .001$) of detecting DIF when the *focal group* came from a lower mean distribution on the first dimension (Mean1 FG=-0.5, Mean2 FG=0), when compared to the MGRM-DFIT method ($\hat{\pi}=.22$).

For item 39, the MG-CFA method ($\hat{\pi}=.26$) had a significantly higher likelihood ($B=.121$, $OR=1.128$, $p < .001$) of detecting DIF when the *focal group* came from a lower mean distribution on the second dimension (Mean1 FG=0, Mean2 FG=-0.5), compared to the MGRM-DFIT method ($\hat{\pi}=.15$). In addition, the MG-CFA method ($\hat{\pi}=.35$) had a significantly higher likelihood ($B=.153$, $OR=1.165$, $p < .001$) of detecting DIF when the *focal group* came from a lower mean distribution on the first dimension (Mean1 FG=-0.5, Mean2 FG=0), compared to the MGRM-DFIT method ($\hat{\pi}=.23$).

For item 40, the MG-CFA method ($\hat{\pi}=.26$) had a significantly higher likelihood ($B=.107$, $OR=1.113$, $p < .001$) of detecting DIF when the *focal group* came from a lower mean distribution on the second dimension (Mean1 FG=0, Mean2 FG=-0.5), compared to the MGRM-DFIT method ($\hat{\pi}=.06$). In addition, the MG-CFA method ($\hat{\pi}=.37$) had a significantly higher likelihood ($B=.171$, $OR=1.186$, $p < .001$) of detecting DIF when the *focal group* came from a lower mean distribution on the first dimension (Mean1 FG=-0.5, Mean2 FG=0), compared to the MGRM-DFIT method ($\hat{\pi}=.15$).

Table 4.2 presents the two-way interaction between latent mean differences and estimation method. This effect was significant for the overall test (all 14 items). When the

focal group came from a lower mean distribution on the second dimension (Mean1 FG=0, Mean2 FG=-0.5), the MGRM-DFIT method ($\hat{\pi}$ =.89) had a significantly higher likelihood (B=.107, OR=1.113, $p < .001$) of detecting DIF when compared to the MG-CFA method ($\hat{\pi}$ =.25). In addition, when the *focal group* came from a lower mean distribution on the first dimension (Mean1 FG=-0.5, Mean2 FG=0), the MGRM-DFIT method ($\hat{\pi}$ =1.00) had a significantly higher likelihood (B=.148, OR=1.159, $p < .001$) of detecting DIF when compared to the MG-CFA method ($\hat{\pi}$ =.40). Finally, when the *focal group* came from a lower mean distribution for both latent dimensions (Mean1 FG=-0.5, Mean2 FG=-0.5), the MGRM-DFIT method ($\hat{\pi}$ =.89) had a significantly higher likelihood (B=.121, OR=1.128, $p < .001$) of detecting DIF compared to the MG-CFA method ($\hat{\pi}$ =.14).

The three-way interaction between uniform DIF, latent mean differences, and estimation method was significant for the overall test (see Table 4.3). For the conditions where it was more difficult for the *reference group* to score in the highest score category (b3=0, b4=-0.5):

- (i) The MGRM-DFIT method ($\hat{\pi}$ =1.00) had a significantly higher likelihood (B=.150, OR=1.162, $p < .001$) of detecting DIF when the *focal group* came from a lower mean distribution on the second dimension (Mean1 FG=0, Mean2 FG=-0.5), when compared to the MG-CFA method ($\hat{\pi}$ =.24).
- (ii) The MGRM-DFIT method ($\hat{\pi}$ =.99) had a significantly higher likelihood (B=.107, OR=1.113, $p < .001$) of detecting DIF when the *focal group* came from a lower mean distribution on the first dimension (Mean1 FG=-0.5, Mean2 FG=0), when compared to the MG-CFA method ($\hat{\pi}$ =.46).

Furthermore, for the conditions where it was more difficult for the *reference group* to score in the two highest score category ($b_3=-0.5$, $b_4=-0.5$):

- (i) The MGRM-DFIT method ($\hat{\pi}=1.00$) had a significantly higher likelihood ($B=.160$, $OR=1.173$, $p < .001$) of detecting DIF when the *focal group* came from a lower mean distribution on the second dimension (Mean1 FG=0, Mean2 FG=-0.5), when compared to the MG-CFA method ($\hat{\pi}=.27$).
- (ii) The MGRM-DFIT method ($\hat{\pi}=.99$) had a significantly higher likelihood ($B=.108$, $OR=1.114$, $p < .001$) of detecting DIF when the *focal group* came from a lower mean distribution on the first dimension (Mean1 FG=-0.5, Mean2 FG=0), when compared to the MG-CFA method ($\hat{\pi}=.58$).

Table 4.4 presents the two-way interaction between uniform DIF and estimation method. This effect was significant for the overall test (all 14 items). When it was more difficult for the *reference group* to score a 5 ($b_3=0$, $b_4=-.5$), the MGRM-DFIT method ($\hat{\pi}=1.00$) had a significantly higher likelihood ($B=.155$, $OR=1.168$, $p < .001$) of detecting DIF compared to the MG-CFA method ($\hat{\pi}=.25$). In addition, the MGRM-DFIT method ($\hat{\pi}=1.00$) had a significantly higher likelihood ($B=.172$, $OR=1.188$, $p < .001$) of detecting DIF when it was more difficult for the *reference group* to score a 4 and a 5 ($b_3=-.5$, $b_4=-.5$), when compared to the MG-CFA method ($\hat{\pi}=.35$).

Table 4.6 presents the proportion of replications, for each estimation method, where DIF was detected for each of the dependent variables averaged across the study conditions. The main effect of estimation method was significant for two of the six dependent variables. The MGRM-DFIT method ($\hat{\pi}=.89$) had a higher likelihood of detecting DIF in the total test (all 14 items) than the MG-CFA method ($\hat{\pi}=.24$), ($B=.976$,

OR=2.655, $p < .001$). However, the MGRM-DFIT method ($\hat{\pi}=.07$) had a lower likelihood of detecting DIF for item 40 (DIF item) than the MG-CFA method ($\hat{\pi}=.16$), (B=-.10, OR=.904, $p < .001$).

A.2 BETWEEN-SUBJECT EFFECTS.

The interaction effect between sample size and latent mean differences was significant for all four DIF items. These results are presented in Table 4.7. For item 37, when the *focal group* came from a lower mean distribution on the second dimension (Mean1 FG=0, Mean2 FG=-0.5), there was a significantly higher likelihood (B=.107, OR=1.112, $p < .001$) of detecting DIF for the N=2000 condition ($\hat{\pi}=.16$), when compared to the N=1000 condition ($\hat{\pi}=.09$). In addition, there was a significantly higher likelihood (B=.113, OR=1.119, $p < .001$) of detecting DIF for the N=2000 condition ($\hat{\pi}=.20$), when the *focal group* came from a lower mean distribution on the first dimension (Mean1 FG=-0.5, Mean2 FG=0), when compared to the N=1000 condition ($\hat{\pi}=.12$).

For item 38, when the *focal group* came from a lower mean distribution on the second dimension (Mean1 FG=0, Mean2 FG=-0.5), there was a significantly higher likelihood (B=.114, OR=1.120, $p < .001$) of detecting DIF for the N=2000 condition ($\hat{\pi}=.20$), when compared to the N=1000 condition ($\hat{\pi}=.13$). In addition, there was a significantly higher likelihood (B=.157, OR=1.170, $p < .001$) of detecting DIF for the N=2000 condition ($\hat{\pi}=.28$), when the *focal group* came from a lower mean distribution on the first dimension (Mean1 FG=-0.5, Mean2 FG=0), when compared to the N=1000 condition ($\hat{\pi}=.17$).

For item 39, when the *focal group* came from a lower mean distribution on the second dimension (Mean1 FG=0, Mean2 FG=-0.5), there was a significantly higher likelihood ($B=.109$, $OR=1.114$, $p < .001$) of detecting DIF for the $N=2000$ condition ($\hat{\pi}=.20$), compared to the $N=1000$ condition ($\hat{\pi}=.13$). In addition, there was a significantly higher likelihood ($B=.146$, $OR=1.157$, $p < .001$) of detecting DIF for the $N=2000$ condition ($\hat{\pi}=.29$), when the *focal group* came from a lower mean distribution on the first dimension (Mean1 FG=-0.5, Mean2 FG=0), when compared to the $N=1000$ condition ($\hat{\pi}=.18$).

Finally, for item 40, when the *focal group* came from a lower mean distribution on the second dimension (Mean1 FG=0, Mean2 FG=-0.5), there was a significantly higher likelihood ($B=.135$, $OR=1.145$, $p < .001$) of detecting DIF for the $N=2000$ condition ($\hat{\pi}=.16$), compared to the $N=1000$ condition ($\hat{\pi}=.09$). In addition, there was a significantly higher likelihood ($B=.147$, $OR=1.158$, $p < .001$) of detecting DIF for the $N=2000$ condition ($\hat{\pi}=.26$), when the *focal group* came from a lower mean distribution on the first dimension (Mean1 FG=-0.5, Mean2 FG=0), when compared to the $N=1000$ condition ($\hat{\pi}=.13$).

The Main effects of Latent mean differences and Uniform DIF were significant for the overall test (all 14 items). The proportion of replications where DIF was detected for each of the six dependent variables across the Latent mean difference and Uniform DIF conditions are presented in Tables 4.8 and 4.9, respectively. For latent mean differences, compared to the *reference* condition ($\hat{\pi}=.48$) with equal latent mean distribution across groups (Mean1 FG=0, Mean2 FG=0), there was a significantly higher likelihood ($B=.114$, $OR=1.121$, $p < .001$) of detecting DIF ($\hat{\pi}=.70$) when the *focal group*

came from a lower mean distribution on the first dimension (Mean1 FG=0, Mean2 FG=-0.5). For uniform DIF, compared to the *reference* condition ($\hat{\pi}=.47$) where it was more difficult for the *focal group* to score a 5 ($b_3=0$, $b_4=+0.5$), there was a significantly higher likelihood ($B=.125$, $OR=1.133$, $p < .001$) of detecting DIF ($\hat{\pi}=.67$) when the items were more difficult for the *reference group* ($b_3=-0.5$, $b_4=-.5$).

APPENDIX B

SAS CODE FOR DATA GENERATION

B.1 TO GENERATE THE BIVARIATE ABILITY DISTRIBUTION

This is the first file to be used for generating the bivariate theta or ability scores for the 'n' examinees. The means for the two dimensions and the variance-covariance matrix has to be specified.

The mvn.sas macro must be saved in the same folder for this to run. The %mvn macro is available from the SAS institute.

The %mvn macro is set-up to run twice for the RG and the FG respectively. sample=theta1 would create the ability scores for 'n' examinees in the RG, and sample=theta2 would generate the ability scores for 'n' examinees in the FG.

"In particular, the means for the RG and FG; 'n' for each group; and the seed numbers need to be specified for each run"

*****/

```
data varcov;  
input v1 v2;  
datalines;  
1 0.6  
0.6 1  
;  
run;
```

```
data means;  
input m;  
datalines;  
0  
0  
;  
run;
```

```
%include 'C:\DIF\mvn.sas';
```

```
%mvn(varcov = varcov, means = means, n = &n1, seed = &seed,  
sample=theta1);
```

```
%mvn(varcov = varcov, means = &meansFG, n = &n2, seed = &seed,  
sample=theta2);
```

```
data theta1;  
set theta1;  
rename col1 = F1 col2 = F2;  
run;
```

```
data theta2;  
set theta2;  
rename col1 = F1 col2 = F2;  
run;
```

B.2 GENERATING MULTIDIMENSIONAL DATA USING PROC IML

```
/******  
Once the theta or ability values are generated using  
gen_bivariate_normal_data.sas, this macro generates the discrete  
scores for 'n' examinees on the 40 items.
```

26 dichotomous items are generated using the MIRT model, and 14 polytomous items are generated using the MGRM.

'a' and 'b' values for the 40 items are generated from a uniform and a normal distribution, respectively.

The start values for the dichotomous and polytomous 'a' values are saved.

This macro creates a '.dat' file with item data for both the RG and the FG for input in Mplus.

In particular, the following need to be specified for each run:

- the magnitude of DIF [&& the direction of DIF (if RG < FG, then magnitude will be specified as a negative number)],
- 'n' for each group,
- the seed number

The following files are created by this macro

- randdich_a_start
- ranpoly_a_start
- group1
- group2
- combined
- combined.dat

```
*****/
```

```
%macro gendata(seed =, n1 =, n2 =, amag1 =, amag2 =, bmag1 =,  
bmag2 =);
```

```
proc iml;
```

```
call randseed(&seed);
```

```
*generating the 'a' values for the 26 dichotomous items from a  
uniform dist;
```

```
aFone13_1 = j(13,1,0);  
call randgen (aFone13_1, 'uniform');  
aFone13_1 = aFone13_1 + 0.75;
```

```
aFone13_2 = j(13,1,0);  
call randgen (aFone13_2, 'normal', 0,0.1);
```

```

aFone13_2 = abs(aFone13_2);

aFone13_2[1,1] = 0;

aFone26 = aFone13_1 // aFone13_2;

aFtwo13_1 = j(13,1,0);
call randgen (aFtwo13_1, 'normal', 0,0.1);
aFtwo13_1 = abs(aFtwo13_1);

aFtwo13_1[1,1] = 0;

aFtwo13_2 = j(13,1,0);
call randgen (aFtwo13_2, 'uniform');
aFtwo13_2 = aFtwo13_2 + 0.75;

aFtwo26 = aFtwo13_1 // aFtwo13_2;

aF26 = aFone26 || aFtwo26;
*to concatenate the A1 and A2 matrices side by side;

dichmaxmajor = max(aF26[1:13,1],aF26[14:26,2]);
dichminmajor = min(aF26[1:13,1],aF26[14:26,2]);

dichmaxminor = max(aF26[2:13,2],aF26[15:26,1]);
dichminminor = min(aF26[2:13,2],aF26[15:26,1]);

randdich_a
dichmaxmajor||dichminmajor||dichmaxminor||dichminminor;

*the maximum and minimum start values of the generated 'a' values
will be recorded for each rep;

create randdich_a_start from randdich_a[colname = {dichmaxmajor
dichminmajor dichmaxminor dichminminor}];
append from randdich_a;
close randdich_a_start;

*generating the 'a' values for the 14 polytomous items from a
uniform dist;

aFone7_1 = j(7,1,0);
call randgen (aFone7_1, 'uniform');
aFone7_1 = aFone7_1 + 0.75;

aFone7_2 = j(7,1,0);
call randgen (aFone7_2, 'normal', 0,0.1);
aFone7_2 = abs(aFone7_2);

aFone7_2[1,1] = 0;

aFone14 = aFone7_1 // aFone7_2;

```

```

aFtwo7_1 = j(7,1,0);
call randgen (aFtwo7_1, 'normal', 0,0.1);
aFtwo7_1 = abs(aFtwo7_1);

aFtwo7_1[1,1] = 0;

aFtwo7_2 = j(7,1,0);
call randgen (aFtwo7_2, 'uniform');
aFtwo7_2 = aFtwo7_2 + 0.75;

aFtwo14 = aFtwo7_1 // aFtwo7_2;

aF14 = aFone14 || aFtwo14;

polymaxmajor = max(aF14[1:7,1],aF14[8:14,2]);
polymminmajor = min(aF14[1:7,1],aF14[8:14,2]);

polymaxminor = max(aF14[2:7,2],aF14[9:14,1]);
polymminminor = min(aF14[2:7,2],aF14[9:14,1]);

randpoly_a =
polymaxmajor||polymminmajor||polymaxminor||polymminminor;

*the maximum and minimum start values of the generated 'a' values
will be recorded for each rep;

create randpoly_a_start from randpoly_a[colname = {polymaxmajor
polymminmajor polymaxminor polymminminor}];
append from randpoly_a;
close randpoly_a_start;

*generating the 'b' values for the 26 dichotomous items from a
normal dist;

b26 = j(26,1,0);
call randgen (b26, 'normal', 0,0.25);

*generating the first 'b' value for the 14 polytomous items from
a normal dist;

b14_1 = j(14,1,0);
call randgen (b14_1, 'normal', -0.5,0.25);

*generating a threshold value that would be added to the
generated 'b1' value to create additional 'b' parameters;

threshadd = j(14,1,0);
call randgen (threshadd, 'uniform');

b14_2 = b14_1 + (threshadd*0.5) + 0.5;

```

```

b14_3 = b14_2 + (threshadd*0.5) + 0.5;

b14_4 = b14_3 + (threshadd*0.5) +0.5;

b14 = b14_1 || b14_2 || b14_3 || b14_4;
*concatenating the 'b1' 'b2' 'b3' and 'b4' values to a matrix;

***** data generation for group 1 (RG) *****;

use thetal;
read all var {f1 f2} into fscoresrg;
close thetal;

psik26 = j(&n1,26,0);
do s = 1 to &n1;
pik26 = j(26,1,0);
do i = 1 to 26;
temp1 = 0;
do h = 1 to 2;
temp1 = (temp1 + (1.7*aF26[i,h])*(fscoresrg[s,h] - b26[i]));
end; *h;
pik26[i] = exp(temp1)/(1+exp(temp1));
u = uniform(0);
if (pik26[i] > u) then psik26[s,i] = 1;
else if (pik26[i] < u) then psik26[s,i] = 0;
end; *i;
end; *s;

rgpsik14 = j(&n1,14,0);
do s = 1 to &n1;
rgpik14 = j(14,4,0);
do i = 1 to 14;
do k = 1 to 4;
temp1 = 0;
do h = 1 to 2;
temp1 = (temp1 + (1.7*aF14[i,h])*(fscoresrg[s,h] - b14[i,k]));
end; *h;
rgpik14[i,k] = exp(temp1)/(1+exp(temp1));
end; *k;
rgpt14 = j(14,5,0);
rgpt14[,1] = 1-rgpik14[,1];
rgpt14[,2] = rgpik14[,1] - rgpik14[,2];
rgpt14[,3] = rgpik14[,2] - rgpik14[,3];
rgpt14[,4] = rgpik14[,3] - rgpik14[,4];
rgpt14[,5] = rgpik14[,4] - 0;
rgcum14 = j(14,5,0);
rgcum14[,1] = rgpt14[,1];
rgcum14[,2] = rgcum14[,1] + rgpt14[,2];
rgcum14[,3] = rgcum14[,2] + rgpt14[,3];
rgcum14[,4] = rgcum14[,3] + rgpt14[,4];
rgcum14[,5] = rgcum14[,4] + rgpt14[,5];
u = uniform(0);

```

```

if (rgcuml4[,1] >= u) then rgpsik14[s,i] = 0;
else if (rgcuml4[,2] >= u) then rgpsik14[s,i] = 1;
else if (rgcuml4[,3] >= u) then rgpsik14[s,i] = 2;
else if (rgcuml4[,4] >= u) then rgpsik14[s,i] = 3;
else if (rgcuml4[,5] >= u) then rgpsik14[s,i] = 4;
end; *i;
end; *s;

rgpsik = psik26 || rgpsik14;

create group1 from rgpsik
[colname = {i1 i2 i3 i4 i5 i6 i7 i8 i9 i10
            i11 i12 i13 i14 i15 i16 i17 i18 i19 i20
            i21 i22 i23 i24 i25 i26 i27 i28 i29 i30
            i31 i32 i33 i34 i35 i36 i37 i38 i39 i40}];

*creates the RG item-data for 'n' students using the parameters
specified;

append from rgpsik;
close;

*****data generation for group 2 (FG)*****;

*the 'a' and 'b' values for the FG are modified to include the
magnitude of DIF for polytomous items 11,12,13&14;

***NOTE: For DIF direction, when RG<FG, then magnitude is
specified as a negative number;

afgF14 = j(14,2,0);
afgF14[1:10,1] = aF14[1:10,1];
afgF14[1:10,2] = aF14[1:10,2];
afgF14[11:14,1] = aF14[11:14,1] + &amag1;
afgF14[11:14,2] = aF14[11:14,2] + &amag2;

bfg14 = j(14,4,0);
bfg14[,1] = b14[,1];
bfg14[,2] = b14[,2];
bfg14[1:10,3] = b14[1:10,3];
bfg14[1:10,4] = b14[1:10,4];
bfg14[11:14,3] = b14[11:14,3] + &bmag1;
bfg14[11:14,4] = b14[11:14,4] + &bmag2;

*data is generated for FG, similar to RG, but using theta2, and
the FG item parms for the polytomous items;
***NOTE: it is important that RG and FG parms be generated within
the same proc iml macro, since the same 'a' and 'b' values are
being used, except for magnitude of DIF in the last 4 items;

use theta2;
read all var {f1 f2} into fscoresfg;

```

```

close thetal1;

psik26 = j(&n2,26,0);
do s = 1 to &n2;
pik26 = j(26,1,0);
do i = 1 to 26;
temp1 = 0;
do h = 1 to 2;
temp1 = (temp1 + (1.7*aF26[i,h])*(fcoresfg[s,h] - b26[i]));
end; *h;
pik26[i] = exp(temp1)/(1+exp(temp1));
u = uniform(0);
if (pik26[i] > u) then psik26[s,i] = 1;
else if (pik26[i] < u) then psik26[s,i] = 0;
end; *i;
end; *s;

fgpsik14 = j(&n2,14,0);
do s = 1 to &n2;
fgpik14 = j(14,4,0);
do i = 1 to 14;
do k = 1 to 4;
temp1 = 0;
do h = 1 to 2;
temp1 = (temp1 + (1.7*afgF14[i,h])*(fcoresfg[s,h] -
bfg14[i,k]));
end; *h;
fgpik14[i,k] = exp(temp1)/(1+exp(temp1));
end; *k;
fgpt14 = j(14,5,0);
fgpt14[,1] = 1-fgpik14[,1];
fgpt14[,2] = fgpik14[,1] - fgpik14[,2];
fgpt14[,3] = fgpik14[,2] - fgpik14[,3];
fgpt14[,4] = fgpik14[,3] - fgpik14[,4];
fgpt14[,5] = fgpik14[,4] - 0;
fgcum14 = j(14,5,0);
fgcum14[,1] = fgpt14[,1];
fgcum14[,2] = fgcum14[,1] + fgpt14[,2];
fgcum14[,3] = fgcum14[,2] + fgpt14[,3];
fgcum14[,4] = fgcum14[,3] + fgpt14[,4];
fgcum14[,5] = fgcum14[,4] + fgpt14[,5];
u = uniform(0);
if (fgcum14[,1] >= u) then fgpsik14[s,i] = 0;
else if (fgcum14[,2] >= u) then fgpsik14[s,i] = 1;
else if (fgcum14[,3] >= u) then fgpsik14[s,i] = 2;
else if (fgcum14[,4] >= u) then fgpsik14[s,i] = 3;
else if (fgcum14[,5] >= u) then fgpsik14[s,i] = 4;
end; *i;
end; *s;

fgpsik = psik26 || fgpsik14;

```



```

create group2 from fgpsik
[colname = {i1 i2 i3 i4 i5 i6 i7 i8 i9 i10
            i11 i12 i13 i14 i15 i16 i17 i18 i19 i20
            i21 i22 i23 i24 i25 i26 i27 i28 i29 i30
            i31 i32 i33 i34 i35 i36 i37 i38 i39 i40}];
append from fgpsik;
close;

*creates the FG itemdata for 'n' students using the parameters
specified;

quit;

%mend gendata;

%gendata(seed=&seed, n1=&n1, n2=&n2, amag1=&amag1, amag2=&amag2,
bmag1=&bmag1, bmag2=&bmag2);

data group1a;
set group1;
group = 1;
run;
data group2a;
set group2;
group = 2;
run;

data combined;
set group1a group2a;
run;

*to create the .dat file to be input in Mplus;

data _null_;
set combined;
file 'c:\dif\combined5.dat';
put (_all_) ('09'X);
run;

*to create the .dat file of the randdich 'a' start values;
data _null_;
set randdich_a_start;
file 'c:\dif\randdich_a_start.dat';
put (_all_) ('09'X);
run;

*to create the .dat file of the randpoly 'a' start values;
data _null_;
set randpoly_a_start;
file 'c:\dif\randpoly_a_start.dat';
put (_all_) ('09'X);
run;

```

APPENDIX C

MPLUS CODES

C.1 BASELINE MODEL

```
title: Project DIF_Dissertation;
data: file is E:\DIF\combined5.dat;
variable: names are i1-i40 group;
usevariables are i1-i40;
categorical i1-i40;
grouping = group (1=rg 2=fg);
analysis: estimator = wlsmv;
model: f1 by i1* i2-i13 i15-i33 i35-i40;
      f2 by i2* i3-i26 i28-i40;
      f1@ 1;
      f2@ 1;
      [f1@0 f2@0];
      f1 with f2;
      {i1-i40@1};
      !specifying scaling factor
model fg: f1 by i1* i2-i13 i15-i33 i35-i40;
      f2 by i2* i3-i26 i28-i40;
      !specifying separate lambda estimates for FG
      f1 with f2;
      !specifying separate correlation estimation for FG
      [i2$1-i13$1 i15$1-i26$1 i27$1-i40$1
      i27$2-i40$2 i27$3-i40$3 i27$4-i40$4];
      !specifying separate threshold estimates for FG
      [f1 f2];
      !specifying separate factor means for FG
savedata: difftest c:\dif\baseline5.dat;
results are C:\DIF\parms.dat;
save = FSCORES;
file is C:\DIF\Fscores.dat;
```

C.2 CONstrained MODEL I (ALL POLYTOMOUS ITEMS CONstrained)

```
title: Project DIF_Dissertation;
data: file is C:\DIF\combined5.dat;
variable: names are i1-i40 group;
usevariables are i1-i40;
categorical i1-i40;
grouping = group (1=rg 2=fg);
analysis: estimator = wlsmv;
difftest = baseline5.dat;
model: f1 by i1* i2-i13 i15-i33 i35-i40;
       f2 by i2* i3-i26 i28-i40;
       f1@ 1;
       f2@ 1;
       [f1@0 f2@0];
       f1 with f2;
       {i1-i40@1};
model fg: f1 by i1* i2-i13 i15-i26;
          f2 by i2* i3-i26;
          f1 with f2;
          [i2$1-i13$1 i15$1-i26$1];
          [f1 f2];
```

C.3 CONSTRAINED MODEL II (10 POLYTOMOUS NON-DIF ITEMS

CONSTRAINED)

```
title: Project DIF_Dissertation;
data: file is C:\DIF\combined5.dat;
variable: names are i1-i40 group;
usevariables are i1-i40;
categorical i1-i40;
grouping = group (1=rg 2=fg);
analysis: estimator = wlsmv;
difftest = baseline5.dat;
model: f1 by i1* i2-i13 i15-i33 i35-i40;
       f2 by i2* i3-i26 i28-i40;
       f1@ 1;
       f2@ 1;
       [f1@0 f2@0];
       f1 with f2;
       {i1-i40@1};
model fg: f1 by i1* i2-i13 i15-i26 i37-i40;
          f2 by i2* i3-i26 i37-i40;
          f1 with f2;
          [i2$1-i13$1 i15$1-i26$1 i37$1-i40$1
           i37$2-i40$2 i37$3-i40$3 i37$4-i40$4];
          [f1 f2];
```

C.4 **CONSTRAINED MODEL III (10 NON-DIF ITEMS + ITEM 37**

CONSTRAINED)

```
title: Project DIF_Dissertation;
data: file is C:\DIF\combined5.dat;
variable: names are i1-i40 group;
usevariables are i1-i40;
categorical i1-i40;
grouping = group (1=rg 2=fg);
analysis: estimator = wlsmv;
difftest = baseline5.dat;
model: f1 by i1* i2-i13 i15-i33 i35-i40;
       f2 by i2* i3-i26 i28-i40;
       f1@ 1;
       f2@ 1;
       [f1@0 f2@0];
       f1 with f2;
       {i1-i40@1};
model fg: f1 by i1* i2-i13 i15-i26 i38-i40;
          f2 by i2* i3-i26 i38-i40;
          f1 with f2;
          [i2$1-i13$1 i15$1-i26$1 i38$1-i40$1
           i38$2-i40$2 i38$3-i40$3 i38$4-i40$4];
          [f1 f2];
```

C.5 CONSTRAINED MODEL IV (10 NON-DIF ITEMS + ITEM 38

CONSTRAINED)

```
title: Project DIF_Dissertation;
data: file is C:\DIF\combined5.dat;
variable: names are i1-i40 group;
usevariables are i1-i40;
categorical i1-i40;
grouping = group (1=rg 2=fg);
analysis: estimator = wlsmv;
difftest = baseline5.dat;
model: f1 by i1* i2-i13 i15-i33 i35-i40;
       f2 by i2* i3-i26 i28-i40;
       f1@ 1;
       f2@ 1;
       [f1@0 f2@0];
       f1 with f2;
       {i1-i40@1};
model fg: f1 by i1* i2-i13 i15-i26 i37 i39-i40;
          f2 by i2* i3-i26 i37 i39-i40;
          f1 with f2;
          [i2$1-i13$1 i15$1-i26$1 i37$1 i39$1 i40$1 i37$2
            i39$2 i40$2 i37$3 i39$3 i40$3 i37$4 i39$4 i40$4];
          [f1 f2];
```

C.6 CONSTRAINED MODEL V (10 NON-DIF ITEMS + ITEM 39

CONSTRAINED)

```
title: Project DIF_Dissertation;
data: file is C:\DIF\combined5.dat;
variable: names are i1-i40 group;
usevariables are i1-i40;
categorical i1-i40;
grouping = group (1=rg 2=fg);
analysis: estimator = wlsmv;
difftest = baseline5.dat;
model: f1 by i1* i2-i13 i15-i33 i35-i40;
      f2 by i2* i3-i26 i28-i40;
      f1@ 1;
      f2@ 1;
      [f1@0 f2@0];
      f1 with f2;
      {i1-i40@1};
model fg: f1 by i1* i2-i13 i15-i26 i37-i38 i40;
      f2 by i2* i3-i26 i37-i38 i40;
      f1 with f2;
      [i2$1-i13$1 i15$1-i26$1 i37$1 i38$1 i40$1
       i37$2 i38$2 i40$2 i37$3 i38$3 i40$3 i37$4
       i38$4 i40$4];
      [f1 f2];
```

C.7 CONstrained Model VI (10 NON-DIF ITEMS + ITEM 40

CONSTRAINED)

```
title: Project DIF_Dissertation;
data: file is C:\DIF\combined5.dat;
variable: names are i1-i40 group;
usevariables are i1-i40;
categorical i1-i40;
grouping = group (1=rg 2=fg);
analysis: estimator = wlsmv;
difftest = baseline5.dat;
model: f1 by i1* i2-i13 i15-i33 i35-i40;
       f2 by i2* i3-i26 i28-i40;
       f1@ 1;
       f2@ 1;
       [f1@0 f2@0];
       f1 with f2;
       {i1-i40@1};
model fg: f1 by i1* i2-i13 i15-i26 i37-i39;
          f2 by i2* i3-i26 i37-i39;
          f1 with f2;
          [i2$1-i13$1 i15$1-i26$1 i37$1-i39$1
           i37$2-i39$2 i37$3-i39$3 i37$4-i39$4];
          [f1 f2];
```


APPENDIX D

SAS CODE FOR READING-IN DATA FROM MPLUS

D.1 READING IN ITEM PARAMETERS GENERATED FROM MPLUS

BASELINE

```
/******  
This file reads the parameter files generated by Mplus for the 5-  
point scale.
```

The parms.dat file contains both the RG and FG parms, but the RG and FG parms are read in separately here.

This file eventually transforms the 'lambda' and 'tau' parameters generated by Mplus into 'a' and 'b' values using proc IML.

Since DIF is estimated only for the polytomous items, only the 'a' and 'b' values for the 14 polytomous items are created here.

**** NO MACRO VARIABLES USED IN THIS FILE...

Creates the following sas data files:

```
- rg_poly  
- fg_poly  
- Chisq_Mplus  
- Chisq_Mplus.dat
```

```
*****/
```

*reading in data for RG, FG parms are dropped here;

```
data rg_parms;  
infile 'c:\dif\parms.dat' lrecl=1000;  
*lrecl indicates the number of characters to read per line;  
input intrg1-intrg82 lamrg1-lamrg76 rhorg i1-i159 x1-x318  
      chisq df p CFI TLI freeparm RMSEA WRMR chisqG1 chisqG2;  
drop i1-i159 x1-x318;  
run;
```

*reading in data for FG, RG parms are dropped here;

```
data fg_parms;  
infile 'c:\dif\parms.dat' lrecl=1000;  
*lrecl indicates the number of characters to read per line;
```

```

input il-il59 intfg1-intfg82 lamfg1-lamfg76 rhofg x1-x318
      chisq df p CFI TLI freeparm RMSEA WRMR chisqG1 chisqG2;
drop il-il59 x1-x318;
run;

*reading in RG parms into proc IML;
proc iml;
use rg_parms;
read all var _num_ into parmsrg;
intrgdich = j(26,1,0);
intrgdich = t(parmsrg[1,1:26]);
intrgpoly = j(56,1,0);
intrgpoly = parmsrg[1,27:82];
intrgpoly = shape(intrgpoly, 14, 4);
lamrgF1dich=
parmsrg[83]//parmsrg[84]//parmsrg[86]//parmsrg[88]//parmsrg[90]//
parmsrg[92]//parmsrg[94]//parmsrg[96]//parmsrg[98]//parmsrg[100]//
/parmsrg[102]//parmsrg[104]//parmsrg[106]//0//parmsrg[109]//parms
rg[111]//parmsrg[113]//parmsrg[115]//parmsrg[117]//parmsrg[119]//
parmsrg[121]//parmsrg[123]//parmsrg[125]//parmsrg[127]//parmsrg[1
29]//parmsrg[131];
lamrgF2dich=
0//parmsrg[85]//parmsrg[87]//parmsrg[89]//parmsrg[91]//parmsrg[93
]//parmsrg[95]//parmsrg[97]//parmsrg[99]//parmsrg[101]//parmsrg[1
03]//parmsrg[105]//parmsrg[107]//parmsrg[108]//parmsrg[110]//parm
srg[112]//parmsrg[114]//parmsrg[116]//parmsrg[118]//parmsrg[120]//
/parmsrg[122]//parmsrg[124]//parmsrg[126]//parmsrg[128]//parmsrg[
130]//parmsrg[132];
lamrgF1poly=
parmsrg[133]//parmsrg[134]//parmsrg[136]//parmsrg[138]//parmsrg[1
40]//parmsrg[142]//parmsrg[144]//0//parmsrg[147]//parmsrg[149]//p
armsrg[151]//parmsrg[153]//parmsrg[155]//parmsrg[157];
lamrgF2poly=
0//parmsrg[135]//parmsrg[137]//parmsrg[139]//parmsrg[141]//parmsr
g[143]//parmsrg[145]//parmsrg[146]//parmsrg[148]//parmsrg[150]//p
armsrg[152]//parmsrg[154]//parmsrg[156]//parmsrg[156];
rhorg = parmsrg[159];

*transforming the 'a' and 'b' values for the polytomous items;

argF1poly = j(14,1,0);
do j = 1 to 14;
argF1poly[j]=(1.7*(lamrgF1poly[j]))/((1-
(lamrgF1poly[j]**2+lamrgF2poly[j]**2+(2*lamrgF1poly[j]*lamrgF2pol
y[j]*rhorg))**.5);
end;

argF2poly = j(14,1,0);
do j = 1 to 14;
argF2poly[j]=(1.7*(lamrgF2poly[j]))/((1-
(lamrgF1poly[j]**2+lamrgF2poly[j]**2+(2*lamrgF1poly[j]*lamrgF2pol
y[j]*rhorg))**.5);

```

```

end;

brgpoly = j(14,4,0);
do j = 1 to 14;
do i= 1 to 4;
brgpoly[j,i]=(intrgpoly[j,i])/((1-
(lamrgF1poly[j]**2+lamrgF2poly[j]**2+(2*lamrgF1poly[j]*lamrgF2pol
y[j]*rhorg))**.5);
end;
end;

argpoly = argF1poly || argF2poly;
rgpoly = argpoly || brgpoly;

itemnum = j(14,1,0);
itemnum = t(1:14);

type = j(14,1,0);
type[1:14,1] = 5;

*creating the RG 'a' and 'b' parameters as a sas data file;
rgpoly = itemnum||type||rgpoly;

create rg_poly from rgpoly [colname = {itemnum type a1 a2 b1 b2
b3 b4}];
append from rgpoly;
close rg_poly;

quit;

*reading in FG parms into proc IML;

proc iml;
use fg_parms;
read all var _num_ into parmsfg;
intfgdich = j(26,1,0);
intfgdich = t(parmsfg[1,1:26]);
intfgpoly = j(56,1,0);
intfgpoly = parmsfg[1,27:82];
intfgpoly = shape(intfgpoly, 14, 4);
lamfgF1dich=
parmsfg[83]//parmsfg[84]//parmsfg[86]//parmsfg[88]//parmsfg[90]//
parmsfg[92]//parmsfg[94]//parmsfg[96]//parmsfg[98]//parmsfg[100]//
/parmsfg[102]//parmsfg[104]//parmsfg[106]//0//parmsfg[109]//parms
fg[111]//parmsfg[113]//parmsfg[115]//parmsfg[117]//parmsfg[119]//
parmsfg[121]//parmsfg[123]//parmsfg[125]//parmsfg[127]//parmsfg[1
29]//parmsfg[131];
lamfgF2dich=
0//parmsfg[85]//parmsfg[87]//parmsfg[89]//parmsfg[91]//parmsfg[93
]//parmsfg[95]//parmsfg[97]//parmsfg[99]//parmsfg[101]//parmsfg[1
03]//parmsfg[105]//parmsfg[107]//parmsfg[108]//parmsfg[110]//parm
sfg[112]//parmsfg[114]//parmsfg[116]//parmsfg[118]//parmsfg[120]//

```

```

/parmsfg[122]//parmsfg[124]//parmsfg[126]//parmsfg[128]//parmsfg[
130]//parmsfg[132];
lamfgF1poly=
parmsfg[133]//parmsfg[134]//parmsfg[136]//parmsfg[138]//parmsfg[1
40]//parmsfg[142]//parmsfg[144]//0//parmsfg[147]//parmsfg[149]//p
armsfg[151]//parmsfg[153]//parmsfg[155]//parmsfg[157];
lamfgF2poly=
0//parmsfg[135]//parmsfg[137]//parmsfg[139]//parmsfg[141]//parmsf
g[143]//parmsfg[145]//parmsfg[146]//parmsfg[148]//parmsfg[150]//p
armsfg[152]//parmsfg[154]//parmsfg[156]//parmsfg[156];
rhofg = parmsfg[159];

*transforming the 'a' and 'b' values for the polytomous items;

afgF1poly = j(14,1,0);
do j = 1 to 14;
afgF1poly[j]=(1.7*(lamfgF1poly[j]))/((1-
(lamfgF1poly[j]**2+lamfgF2poly[j]**2+(2*lamfgF1poly[j]*lamfgF2pol
y[j]*rhofg))**.5);
end;

afgF2poly = j(14,1,0);
do j = 1 to 14;
afgF2poly[j]=(1.7*(lamfgF2poly[j]))/((1-
(lamfgF1poly[j]**2+lamfgF2poly[j]**2+(2*lamfgF1poly[j]*lamfgF2pol
y[j]*rhofg))**.5);
end;

bfgpoly = j(14,4,0);
do j = 1 to 14;
do i= 1 to 4;
bfgpoly[j,i]=(intfgpoly[j,i])/((1-
(lamfgF1poly[j]**2+lamfgF2poly[j]**2+(2*lamfgF1poly[j]*lamfgF2pol
y[j]*rhofg))**.5);
end;
end;

afgpoly = afgF1poly || afgF2poly;
fgpoly = afgpoly || bfgpoly;

itemnum = j(14,1,0);
itemnum = t(1:14);

type = j(14,1,0);
type[1:14,1] = 5;

*creating the FG 'a' and 'b' parameters as a sas data file;
fgpoly = itemnum||type||fgpoly;

create fg_poly from fgpoly [colname = {itemnum type a1 a2 b1 b2
b3 b4}];
append from fgpoly;

```

```

close fg_poly;
quit;

*reading chi-square values for each condition;
data chisq_mplus;
infile 'c:\dif\parms.dat' lrecl=1000; *lrecl indicates the number
of characters to force read per line;
input i1-i318 x1-x318
      chisq df p CFI TLI freeparm RMSEA WRMR chisqG1 chisqG2;
drop i1-i318 x1-x318;
run;

*to create the .dat file of the NCDIF for each item for each
replication;
data _null_;
set chisq_mplus;
file 'c:\dif\chisq_mplus.dat';
put (_all_) ('09'X);
run;

```

D.2 READING IN ABILITY ESTIMATES GENERATED FROM MPLUS

BASELINE

```
/******  
This file reads the Fscores files generated by Mplus, and creates  
sas data files with generated theta scores.
```

```
Specifically the theta scores for the polytomous items are  
created for use by the MGRM-DFIT macro.
```

```
**** NO MACRO VARIABLES USED IN THIS FILE...
```

```
Creates the following sas data files:
```

```
- fscores_rg  
- fscores_fg  
- itemscores_rg  
- itemscores_fg  
- polyitemscores_rg
```

```
*****/
```

```
data fscores;  
infile 'c:\dif\fscores.dat' lrecl=1000;  
*lrecl indicates the number of characters to read per line;  
input il-i40 f1 f2 group;  
run;
```

```
data fscores1;  
set fscores;  
keep f1 f2 group;  
run;
```

```
data fscores_rg fscores_fg;  
set fscores1;  
select (group);  
when ('1') output fscores_rg;  
when ('2') output fscores_fg;  
otherwise;  
end;  
drop group;  
run;
```

```
data itemscores;  
set fscores;  
keep il-i40 group;  
run;
```

```
data itemscores_rg itemscores_fg;  
set itemscores;  
select (group);
```

```
when ('1') output itemscores_rg;  
when ('2') output itemscores_fg;  
    otherwise;  
    end;  
drop group;  
run;
```

```
data polyitemscores_rg;  
set itemscores_rg;  
drop i1-i26;  
run;
```

```
data polyitemscores_rg;  
set polyitemscores_rg;  
rename i27-i40 = i1-i14;  
run;
```

D.3 READING IN CHI-SQUARE DIFFERENCE TESTS FROM MPLUS

CONSTRAINED MODELS

```
/******  
This file reads the all the 6 chisq difference tests from the  
Mplus output, and creates sas data files with chisq, df, and p  
values. An appended .dat file is created with all 6 chisq  
difference tests.
```

```
**** MACRO VARIABLE USED IN THIS FILE:  
- &cond (no need to specify)
```

```
Creates the following sas data files:
```

```
- Chisq_diff51  
- Chisq_diff52  
- Chisq_diff53  
- Chisq_diff54  
- Chisq_diff55  
- Chisq_diff56  
- Chisq_diff.dat
```

```
*****/
```

```
*DATA JUNK IS CREATED TO INPUT MISSING VALUES IN CASE ONE OF THE  
MEASUREMENT INVARIANCE MODELS HAS NOT CONVERGED;
```

```
data junk;  
input chisq;  
datalines;  
-999  
-999  
-999  
;  
run;
```

```
*READING IN THE CHI-SQUARE DIFFERENCE TEST FOR THE FIRST  
CONSTRAINED MODEL;
```

```
data chisq_diff51;  
infile 'c:\dif\dif_measurementinvariance_51.out' lrecl=1000;  
*lrecl indicates the number of characters to force read per line;  
input string $ 1-50;  
if (string = "Chi-Square Test for Difference Testing") then do;  
do i = 1 to 1;  
input;  
end; *do i;  
do j = 1 to 3;  
input chisq 42-50;  
output;  
end; *do j;
```



```

end; *if;
drop i j string;
run;

data chisq_diff51;
set chisq_diff51 junk;
if _N_ >= 4 then delete;
run;

data chisq_diff;
set chisq_diff51;
run;

*READING IN THE CHI-SQUARE DIFFERENCE TEST FOR THE SECOND
CONSTRAINED MODEL;

data chisq_diff52;
infile 'c:\dif\dif_measurementinvariance_52.out' lrecl=1000;
*lrecl indicates the number of characters to read per line;
input string $ 1-50;
if (string = "Chi-Square Test for Difference Testing") then do;
do i = 1 to 1;
input;
end; *do i;
do j = 1 to 3;
input chisq 42-50;
output;
end; *do j;
end; *if;
drop i j string;
run;

data chisq_diff52;
set chisq_diff52 junk;
if _N_ >= 4 then delete;
run;

* APPENDING CHI-SQUARE DIFFERENCE TESTS TO THE SAME FILE;

data chisq_diff;
set chisq_diff;

proc append BASE = chisq_diff DATA=chisq_diff52;
run;

*READING IN THE CHI-SQUARE DIFFERENCE TEST FOR THE THIRD
CONSTRAINED MODEL;

data chisq_diff53;
infile 'c:\dif\dif_measurementinvariance_53.out' lrecl=1000;
*lrecl indicates the number of characters to read per line;
input string $ 1-50;

```

```

if (string = "Chi-Square Test for Difference Testing") then do;
do i = 1 to 1;
input;
end; *do i;
do j = 1 to 3;
input chisq 42-50;
output;
end; *do j;
end; *if;
drop i j string;
run;

data chisq_diff53;
set chisq_diff53 junk;
if _N_ >= 4 then delete;
run;

* APPENDING CHI-SQUARE DIFFERENCE TESTS TO THE SAME FILE;

data chisq_diff;
set chisq_diff;

proc append BASE = chisq_diff DATA=chisq_diff53;
run;

*READING IN THE CHI-SQUARE DIFFERENCE TEST FOR THE FOURTH
CONSTRAINED MODEL;

data chisq_diff54;
infile 'c:\dif\dif_measurementinvariance_54.out' lrecl=1000;
*lrecl indicates the number of characters to read per line;
input string $ 1-50;
if (string = "Chi-Square Test for Difference Testing") then do;
do i = 1 to 1;
input;
end; *do i;
do j = 1 to 3;
input chisq 42-50;
output;
end; *do j;
end; *if;
drop i j string;
run;

data chisq_diff54;
set chisq_diff54 junk;
if _N_ >= 4 then delete;
run;

* APPENDING CHI-SQUARE DIFFERENCE TESTS TO THE SAME FILE;

data chisq_diff;

```

```

set chisq_diff;

proc append BASE = chisq_diff DATA=chisq_diff54;
run;

*READING IN THE CHI-SQUARE DIFFERENCE TEST FOR THE FIFTH
CONSTRAINED MODEL;

data chisq_diff55;
infile 'c:\dif\dif_measurementinvariance_55.out' lrecl=1000;
*lrecl indicates the number of characters to read per line;
input string $ 1-50;
if (string = "Chi-Square Test for Difference Testing") then do;
do i = 1 to 1;
input;
end; *do i;
do j = 1 to 3;
input chisq 42-50;
output;
end; *do j;
end; *if;
drop i j string;
run;

data chisq_diff55;
set chisq_diff55 junk;
if _N_ >= 4 then delete;
run;

* APPENDING CHI-SQUARE DIFFERENCE TESTS TO THE SAME FILE;

data chisq_diff;
set chisq_diff;

proc append BASE = chisq_diff DATA=chisq_diff55;
run;

*READING IN THE CHI-SQUARE DIFFERENCE TEST FOR THE SIXTH
CONSTRAINED MODEL;

data chisq_diff56;
infile 'c:\dif\dif_measurementinvariance_56.out' lrecl=1000;
*lrecl indicates the number of characters to read per line;
input string $ 1-50;
if (string = "Chi-Square Test for Difference Testing") then do;
do i = 1 to 1;
input;
end; *do i;
do j = 1 to 3;
input chisq 42-50;
output;
end; *do j;

```

```

end; *if;
drop i j string;
run;

data chisq_diff56;
set chisq_diff56 junk;
if _N_ >= 4 then delete;
run;

* APPENDING CHI-SQUARE DIFFERENCE TESTS TO THE SAME FILE;

data chisq_diff;
set chisq_diff;

proc append BASE = chisq_diff DATA=chisq_diff56;
run;

*to create the .dat file of the chisq_diff values from Mplus;
data _null_;
set chisq_diff;
file 'c:\dif\chisq_diff.dat';
put (_all_) ('09'X);
run;

```

APPENDIX E

SAS CODES FOR RUNNING THE MGRM-DFIT MACROS

E.1 CREATING DATASETS FOR DFIT CHISQUARE TESTING

*****;
This step creates the FG, RG item parameter datasets required for each DFIT chisquare testing (with items removed).

This step is to be performed only the linked FG parms are read in from LinkMIRT and fgpoly_linked is created.

This step also creates the polytomous itemdata for RG with each item removed.

*** NO MACRO VARIABLES ARE USED IN THIS FILE...

Creates the following sas data files:

```
- fg_poly_no_DIF; fg_poly_no_DIF_11 ... .. fg_poly_no_DIF_14
- rg_poly_no_DIF; rg_poly_no_DIF_11 ... .. rg_poly_no_DIF_14
- polyitemscores_rg_no_DIF;
- polyitemscores_rg_no_DIF_11 ... .. polyitemscores_rg_no_DIF_14
*****;
```

```
data rg_poly_no_DIF;
set rg_poly;
if itemnum notin (1,2,3,4,5,6,7,8,9,10) then delete;
run;
```

```
data rg_poly_no_DIF_11;
set rg_poly;
if itemnum notin (1,2,3,4,5,6,7,8,9,10,11) then delete;
run;
```

```
data rg_poly_no_DIF_12;
set rg_poly;
if itemnum notin (1,2,3,4,5,6,7,8,9,10,12) then delete;
run;
```

```
data rg_poly_no_DIF_13;
set rg_poly;
if itemnum notin (1,2,3,4,5,6,7,8,9,10,13) then delete;
run;
```

```

data rg_poly_no_DIF_14;
set rg_poly;
if itemnum notin (1,2,3,4,5,6,7,8,9,10,14) then delete;
run;

*****;

data fg_poly_no_DIF;
set fg_poly;
if itemnum notin (1,2,3,4,5,6,7,8,9,10) then delete;
run;

data fg_poly_no_DIF_11;
set fg_poly;
if itemnum notin (1,2,3,4,5,6,7,8,9,10,11) then delete;
run;

data fg_poly_no_DIF_12;
set fg_poly;
if itemnum notin (1,2,3,4,5,6,7,8,9,10,12) then delete;
run;

data fg_poly_no_DIF_13;
set fg_poly;
if itemnum notin (1,2,3,4,5,6,7,8,9,10,13) then delete;
run;

data fg_poly_no_DIF_14;
set fg_poly;
if itemnum notin (1,2,3,4,5,6,7,8,9,10,14) then delete;
run;

*****;

data polyitemscores_rg_no_DIF;
set polyitemscores_rg;
drop i11 i12 i13 i14;
run;

data polyitemscores_rg_no_DIF_11;
set polyitemscores_rg;
drop i12 i13 i14;
run;

data polyitemscores_rg_no_DIF_12;
set polyitemscores_rg;
drop i11 i13 i14;
rename i12 = i11;
run;

data polyitemscores_rg_no_DIF_13;
set polyitemscores_rg;

```

```
drop i11 i12 i14;  
rename i13 = i11;  
run;  
  
data polyitemscores_rg_no_DIF_14;  
set polyitemscores_rg;  
drop i11 i12 i13;  
rename i14 = i11;  
run;
```

E.2 MGRM-DFIT FOR ALL 14 POLYTOMOUS ITEMS

```

/*****
This step performs the MGRM DFIT on the 14 polytomous items using
proc IML.
THE FOLLOWING MACRO VARIABLES NEED TO BE SPECIFIED HERE:
- nl for group 1 (since RG item scores are used for both RG and
FG computations of ESSi) only the 'n' for the RG will be used for
DFIT computations...
THE FOLLOWING MACRO VARIABLES WILL NOT CHANGE FROM RUN TO RUN IN
THIS FILE:
- fg_poly (since all items are used for the test here, the same
file will always be used)
- rg_poly (again, since all items are used, this file will not
change)
- polyitemscores_rg (again, since all items are used, this file
will not change)
- i2 - refers to the total number of items for this DFIT chisq
testing, and it will refer to all 14 items here
- btotat - refers to the total number of 'b' parameters - for a
5-point scale, this would refer to 4*14 = 56
- cumbtotat - refers to the total number of categories - for a 5-
point scale, this would be 5*14 = 70
THE FOLLOWING SAS DATA FILES ARE CREATED IN THIS STEP:
- ES_Ref_rg - required for CDIF & NCDIF testing
- ES_Ref_fg - required for CDIF & NCDIF testing
- T_Ref_rg - required for DTF testing
- T_Ref_fg - required for DTF testing
- T_Ref - combined file of T_Ref_rg and T_Ref_fg used for
computing DsquareS
- D - file need computing NCDIF and CDIF
- MeanD - needed for chisquare testing
- chisqDTF - a file that contains the chisq test for all items,
and will be appended to, after the chisq tests, in the following
steps
*****/

%macro dfit (nl= , fg_poly= , rg_poly= , polyitemscores_rg= , i2=
, btotat= , cumbtotat=);

proc iml;
use &rg_poly; *the rg item parameters for all 14 polytomous
items;
read all var ('A1':'A2') into arg;
read all var ('B1':'B4') into brg;
close &rg_poly;

use fscores_rg; *the theta estimates for rg from Mplus;
read all var {f1 f2} into fscores;
close fscores_rg;

```



```

psik = j(&n1,&bttotal,0);

do s = 1 to &n1;
pik = j(&i2,4,0);
do i = 1 to &i2;
do k = 1 to 4;
temp1 = 0;
do h = 1 to 2;
temp1 = (temp1 + arg[i,h]*(fscores[s,h] - brg[i,k]));
*the MGRM psik is computed using computed parms from the Mplus
baseline model;
end; *h;
pik[i,k] = exp(temp1)/(1+exp(temp1));
end; *k;
end; *i;
psik[s,] = shape(pik,1,&bttotal);
end; *s;

ptheta = j(&n1,&cumbtotal,0);

do s = 1 to &n1;
do i = 1 to &i2;
do j = 1 to 5;
temp2 = (i-1)*5+j;
temp3 = (i-1)*4+j;

*discrete item scores are computed below;

if (j = 1) then ptheta[s,temp2] = 1 - psik[s,temp3];
else if (j >= 2 & j <= 4) then ptheta[s,temp2] = psik[s,temp3-1]
- psik[s,temp3];
else if (j = 5) then ptheta[s,temp2] = psik[s,temp3-1];
end;
end;
end;

use &polyitemscores_rg;
*using the item scores for RG generated from Mplus;
read all var ('i1':'i14') into itemscores;
close &polyitemscores_rg;

*ESSi for each subject (sum of Psik for each item) is computed
below;
*also the total score for each subject is computed below;

itemscores = itemscores + 1;

ESRrg = j(&n1,&i2,0);
TRrg = j(&n1,1,0);
do s = 1 to &n1;
do i = 1 to &i2;
ESRrg[s,i] = ptheta[s,(i-1)*5+(itemscores[s,i])]*itemscores[s,i];

```

```

end;
TRrg[s] = sum(ESRrg[s,1:&i2]);
end;

create ES_Ref_rg from ESRrg[colname = {i1 i2 i3 i4 i5 i6 i7 i8 i9
i10 i11 i12 i13 i14}];
append from ESRrg;
close ES_Ref_rg;

Truescore = {TRefRG};
create T_Ref_rg from TRrg[COLNAME=Truescore];
append from TRrg;
close T_Ref_rg;

quit;

proc iml;
use &fg_poly; *the fg item parameters for all 14 polytomous
items;
read all var ('A1':'A2') into afg;
read all var ('B1':'B4') into bfg;
close &fg_poly;

use fscores_rg; *the theta estimates for rg from Mplus;
read all var {f1 f2} into fscores;
close fscores_rg;

psik = j(&n1,&bttotal,0);

do s = 1 to &n1;
pik = j(&i2,4,0);
do i = 1 to &i2;
do k = 1 to 4;
temp1 = 0;
do h = 1 to 2;
temp1 = (temp1 + afg[i,h]*(fscores[s,h] - bfg[i,k]));
*the MGRM psik is computed using computed parms from the Mplus
baseline model;
end; *h;
pik[i,k] = exp(temp1)/(1+exp(temp1));
end; *k;
end; *i;
psik[s,] = shape(pik,1,&bttotal);
end; *s;

ptheta = j(&n1,&cumbtotal,0);

do s = 1 to &n1;
do i = 1 to &i2;
do j = 1 to 5;
temp2 = (i-1)*5+j;
temp3 = (i-1)*4+j;

```

```

*discrete item scores are computed below;

if (j = 1) then ptheta[s,temp2] = 1 - psik[s,temp3];
else if (j >= 2 & j <= 4) then ptheta[s,temp2] = psik[s,temp3-1]
- psik[s,temp3];
else if (j = 5) then ptheta[s,temp2] = psik[s,temp3-1];

end;
end;
end;

use polyitemscores_rg;
*using the item scores for RG generated from Mplus;
read all var ('i1':'i14') into itemscores;
close polyitemscores_rg;

*ESSi for each subject (sum of Psik for each item) is computed
below;
*also the total score for each subject is computed below;

itemscores = itemscores + 1;

ESRfg = j(&n1,&i2,0);
TRfg = j(&n1,1,0);
do s = 1 to &n1;
do i = 1 to &i2;
ESRfg[s,i] = ptheta[s,(i-1)*5+(itemscores[s,i])]*itemscores[s,i];
end;
TRfg[s] = sum(ESRfg[s,1:&i2]);
end;

create ES_Ref_fg from ESRfg[colname = {i1 i2 i3 i4 i5 i6 i7 i8 i9
i10 i11 i12 i13 i14}];
append from ESRfg;
close ES_Ref_fg;

Truescore = {TRefFG};
create T_Ref_fg from TRfg[COLNAME=Truescore];
append from TRfg;
close T_Ref_fg;

quit;

*the total test score for RG and FG are used to compute DTF;

data T_Ref_rg;
set T_Ref_rg;
Subject = _N_;
run;

data T_Ref_fg;
set T_Ref_fg;

```

```

Subject = _N_;
run;

data T_Ref;
merge T_Ref_rg T_Ref_fg;
by Subject;
drop subject;
run;

data D;
set T_Ref;
D = (TRefRG-TRefFG)**2;
*D = DsquareS**.5;
run;

proc means noprint data = D;
var TRefRG TRefFG D;
output out =
MeanD MEAN(TRefRG TRefFG D) = MeanTRefRG MeanTRefFG MeanD
STDDEV(TRefRG TRefFG D) = StddevTRefRG StddevTRefFG StddevD;
run;

data MeanD;
set MeanD;
keep MeanTRefRG MeanTRefFG MeanD StddevTRefRG StddevTRefFG
StddevD;
run;

data Mean1;
set MeanD;
keep StddevD;
run;

proc iml;
use MeanD;
read all var {MeanTRefRG MeanTRefFG MeanD StddevTRefRG
StddevTRefFG StddevD} into MDTF;
close MeanD;

use Mean1;
read all var {StddevD} into VARD;
close Mean1;

DTF = j(1,1,0);
DTF = MDTF[,6]**2 + MDTF[,3]**2;

chil = DTF*&n1;
chisq = chil / (VARD[,1]**2);
p = 1 - (probchi(chisq,&n1));

chiDTF = DTF||chisq||p;
create chisqDTF from chiDTF[colname = {DTF chisq p}];

```

```

append from chiDTF;
close chisqDTF;

quit;

%mend dfit;

%dfit (n1 = &n1, fg_poly = fg_poly, rg_poly = rg_poly,
polyitemscores_rg = polyitemscores_rg, i2 = 14, btotal = 56,
cumbtotal = 70);

```

E.3 MGRM-DFIT COMPUTING CDIF AND NCDIF FOR ALL 14 ITEMS

```

/*****
This step performs the CDIF and NCDIF item level tests for the 14
polytomous items.
THE FOLLOWING SAS DATA FILES ARE REQUIRED IN THIS STEP:
- ES_Ref_rg - required for CDIF & NCDIF testing
- ES_Ref_fg - required for CDIF & NCDIF testing
- D - file needed for computing NCDIF and CDIF
THE FOLLOWING SAS DATA & .dat FILES ARE CREATED IN THIS STEP:
- D_i
- MeanD_i
- NCDIFi
- NCDIFi.dat
- CDIF
- covdi
- CDIFi
- CDIFi.dat
**** MACRO VARIABLE USED IN THIS FILE:
- &n1 (The RG 'n' value will be used here
*****/

*READING IN DATA FROM THE DFIT RUN FOR ALL ITEMS;

proc iml;
use ES_Ref_rg;
read all var ('i1':'i14') into dirg;
close ES_Ref_rg;

use ES_Ref_fg;
read all var ('i1':'i14') into difg;
close ES_Ref_fg;

di = j(&n1,14,0);
do s = 1 to &n1;
do i = 1 to 14;
di[s,i] = dirg[s,i] - difg[s,i];
end;
end;

create d_i from di[colname = {i1 i2 i3 i4 i5 i6 i7 i8 i9 i10 i11
i12 i13 i14}];
append from di;
close d_i;

quit;

proc means noprint data = D_i;
var i1-i14;
output out =
```

```

MeanD_i MEAN(i1 i2 i3 i4 i5 i6 i7 i8 i9 i10 i11 i12 i13 i14) =
Meani1 Meani2 Meani3 Meani4 Meani5 Meani6 Meani7
Meani8 Meani9 Meani10 Meani11 Meani12 Meani13 Meani14
STDDEV(i1 i2 i3 i4 i5 i6 i7 i8 i9 i10 i11 i12 i13 i14) =
Stddevi1 Stddevi2 Stddevi3 Stddevi4 Stddevi5 Stddevi6 Stddevi7
Stddevi8 Stddevi9 Stddevi10 Stddevi11 Stddevi12 Stddevi13
Stddevi14;
run;

data MeanD_i;
set MeanD_i;
keep Meani1 Meani2 Meani3 Meani4 Meani5 Meani6 Meani7
Meani8 Meani9 Meani10 Meani11 Meani12 Meani13 Meani14
Stddevi1 Stddevi2 Stddevi3 Stddevi4 Stddevi5 Stddevi6 Stddevi7
Stddevi8 Stddevi9 Stddevi10 Stddevi11 Stddevi12 Stddevi13
Stddevi14;
run;

proc iml;
use MeanD_i;
read all var {Meani1 Meani2 Meani3 Meani4 Meani5 Meani6 Meani7
Meani8 Meani9 Meani10 Meani11 Meani12 Meani13 Meani14} into
Mean_di;
read all var {Stddevi1 Stddevi2 Stddevi3 Stddevi4 Stddevi5
Stddevi6 Stddevi7 Stddevi8 Stddevi9 Stddevi10 Stddevi11 Stddevi12
Stddevi13 Stddevi14} into SD_di;
close MeanD_i;

Mean_di = t(Mean_di);
SD_di = t(SD_di);

*COMPUTING NCDIF FOR ALL 14 ITEMS;

NCDIF = j(14,1,0);
do i = 1 to 14;
NCDIF[i] = Mean_di[i]**2 + SD_di[i]**2;
end;

NCDIF = t(NCDIF);

create NCDIFi from NCDIF[colname = {i1 i2 i3 i4 i5 i6 i7 i8 i9
i10 i11 i12 i13 i14}];
append from NCDIF;
close NCDIFi;

quit;

*to create the .dat file of the NCDIF for each item for each
replication;
data _null_;
set NCDIFi;
file 'c:\dif\NCDIFi.dat';

```

```

put (_all_) ('09'X);
run;

*COMPUTING CDIF FOR ALL 14 ITEMS;

data D_i;
set D_i;
s = _N_;
run;

data D;
set D;
s = _N_;
run;

data CDIF;
merge D_i D;
by s;
run;

proc corr noprint data = cdif outp = covdi cov;
var i1-i14;
run;

data covdi;
set covdi;
n = _N_;
run;

data covdi;
set covdi;
if n notin (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14) then
delete;
keep i1-i14;
run;

proc iml;
use covdi;
read all var ('i1':'i14') into Covdi;
close covdi;

use MeanD_i;
read all var {Meani1 Meani2 Meani3 Meani4 Meani5 Meani6 Meani7
Meani8 Meani9 Meani10 Meani11 Meani12 Meani13 Meani14} into
Mean_di;
read all var {Stddevi1 Stddevi2 Stddevi3 Stddevi4 Stddevi5
Stddevi6 Stddevi7 Stddevi8 Stddevi9 Stddevi10 Stddevi11 Stddevi12
Stddevi13 Stddevi14} into SD_di;
close MeanD_i;

Mean_di = t(Mean_di);
SD_di = t(SD_di);

```



```

var_di = j(14,1,0);
do i = 1 to 14;
var_di[i] = SD_di[i]**2;
end;

use MeanD;
read all var {MeanD StddevD} into MD;
close MeanD;

covdi1 = sum(Covdi[1,2:14]);
covdiD1 = covdi1 + var_di[1,1];
CDIFi1 = covdiD1 +(Mean_di[1,1]*MD[1,1]);

covdi2 = sum(Covdi[2,3:14]);
covdi2 = covdi2 + Covdi[2,1];
covdiD2 = covdi2 + var_di[2,1];
CDIFi2 = covdiD2 +(Mean_di[2,1]*MD[1,1]);

covdi31 = sum(Covdi[3,1:2]);
covdi32 = sum(Covdi[3,4:14]);
covdi3 = covdi31 + covdi32;
covdiD3 = covdi3 + var_di[3,1];
CDIFi3 = covdiD3 +(Mean_di[3,1]*MD[1,1]);

covdi41 = sum(Covdi[4,1:3]);
covdi42 = sum(Covdi[4,5:14]);
covdi4 = covdi41 + covdi42;
covdiD4 = covdi4 + var_di[4,1];
CDIFi4 = covdiD4 +(Mean_di[4,1]*MD[1,1]);

covdi51 = sum(Covdi[5,1:4]);
covdi52 = sum(Covdi[5,6:14]);
covdi5 = covdi51 + covdi52;
covdiD5 = covdi5 + var_di[5,1];
CDIFi5 = covdiD5 +(Mean_di[5,1]*MD[1,1]);

covdi61 = sum(Covdi[6,1:5]);
covdi62 = sum(Covdi[6,7:14]);
covdi6 = covdi61 + covdi62;
covdiD6 = covdi6 + var_di[6,1];
CDIFi6 = covdiD6 +(Mean_di[6,1]*MD[1,1]);

covdi71 = sum(Covdi[7,1:6]);
covdi72 = sum(Covdi[7,8:14]);
covdi7 = covdi71 + covdi72;
covdiD7 = covdi7 + var_di[7,1];
CDIFi7 = covdiD7 +(Mean_di[7,1]*MD[1,1]);

covdi81 = sum(Covdi[8,1:7]);
covdi82 = sum(Covdi[8,9:14]);
covdi8 = covdi81 + covdi82;
covdiD8 = covdi8 + var_di[8,1];

```

```

CDIFi8 = covdiD8 +(Mean_di[8,1]*MD[1,1]);

covdi91 = sum(Covdi[9,1:8]);
covdi92 = sum(Covdi[9,10:14]);
covdi9 = covdi91 + covdi92;
covdiD9 = covdi9 + var_di[9,1];
CDIFi9 = covdiD9 +(Mean_di[9,1]*MD[1,1]);

covdi101 = sum(Covdi[10,1:9]);
covdi102 = sum(Covdi[10,11:14]);
covdi10 = covdi101 + covdi102;
covdiD10 = covdi10 + var_di[10,1];
CDIFi10 = covdiD10 +(Mean_di[10,1]*MD[1,1]);

covdi111 = sum(Covdi[11,1:10]);
covdi112 = sum(Covdi[11,12:14]);
covdi11 = covdi111 + covdi112;
covdiD11 = covdi11 + var_di[11,1];
CDIFi11 = covdiD11 +(Mean_di[11,1]*MD[1,1]);

covdi121 = sum(Covdi[12,1:11]);
covdi122 = sum(Covdi[12,13:14]);
covdi12 = covdi121 + covdi122;
covdiD12 = covdi12 + var_di[12,1];
CDIFi12 = covdiD12 +(Mean_di[12,1]*MD[1,1]);

covdi13 = sum(Covdi[13,1:12]);
covdi13 = covdi13 + Covdi[13,14];
covdiD13 = covdi13 + var_di[13,1];
CDIFi13 = covdiD13 +(Mean_di[13,1]*MD[1,1]);

covdi14 = sum(Covdi[14,1:13]);
covdiD14 = covdi14 + var_di[14,1];
CDIFi14 = covdiD14 +(Mean_di[14,1]*MD[1,1]);

CDIFi = CDIFi1||CDIFi2||CDIFi3||CDIFi4||CDIFi5||CDIFi6||CDIFi7||
CDIFi8||CDIFi9||CDIFi10||CDIFi11||CDIFi12||CDIFi13||CDIFi14;
create CDIFi from CDIFi
[colname = {i1 i2 i3 i4 i5 i6 i7 i8 i9 i10 i11 i12 i13 i14}];
append from CDIFi;
close CDIFi;
quit;

*to create the .dat file of the CDIF for each item for each
replication;
data _null_;
set CDIFi;
file 'c:\dif\CDIFi.dat';
put (_all_) ('09'X);
run;

```

E.4 MGRM-DFIT FOR THE 10 NON-DIF POLYTOMOUS ITEMS

```

/*****
This step performs the MGRM DFIT on the 10 polytomous items with
no DIF, using proc IML.
THE FOLLOWING MACRO VARIABLES NEED TO BE SPECIFIED HERE:
- nl for group 1 (since RG item scores are used for both RG and
FG computations of ESSi) only the 'n' for the RG will be used for
DFIT computations...
THE FOLLOWING MACRO VARIABLES WILL NOT CHANGE FROM RUN TO RUN IN
THIS FILE:
- fg_poly_no_DIF (this includes the FG polytomous item parameters
for the 10 non-DIF items)
- rg_poly_no_DIF (this includes the RG polytomous item parameters
for the 10 non-DIF items)
- i2 - refers to the total number of items for this DFIT chisq
testing, and it will refer to 10 items here
- btotal - refers to the total number of 'b' parameters - for a
5-point scale, this would refer to 4*10 = 40
- cumbtotal - refers to the total number of categories - for a 5-
point scale, this would be 5*10 = 50
THE FOLLOWING SAS DATA FILES ARE CREATED IN THIS STEP:
- ES_Ref_rg - required for CDIF & NCDIF testing
- ES_Ref_fg - required for CDIF & NCDIF testing
- T_Ref_rg - required for DTF testing
- T_Ref_fg - required for DTF testing
- T_Ref - combined file of T_Ref_rg and T_Ref_fg used for
computing DsquareS
- D - file need computing NCDIF and CDIF
- MeanD - needed for chisquare testing
- chisqDTF1 - a file that contains the chisq test for the
particular DFIT test
- chisqDTF - a file that contains the appended chisq tests for
all DFIT tests, (this file will be appended to, after each chisq
test, in this and the following steps)
*****/

%macro dfit (nl= , fg_poly= , rg_poly= , polyitemscores_rg= , i2=
, btotal= , cumbtotal=);

proc iml;
use &rg_poly;
*the rg item parameters for the 10 non-dif polytomous items;
read all var ('A1':'A2') into arg;
read all var ('B1':'B4') into brg;
close &rg_poly;

use fscores_rg; *the theta estimates for rg from mplus;
read all var {f1 f2} into fscores;
close fscores_rg;

```

```

psik = j(&n1,&bttotal,0);

do s = 1 to &n1;
pik = j(&i2,4,0);
do i = 1 to &i2;
do k = 1 to 4;
temp1 = 0;
do h = 1 to 2;
temp1 = (temp1 + arg[i,h]*(fscores[s,h] - brg[i,k]));
parms;
end; *h;
pik[i,k] = exp(temp1)/(1+exp(temp1));
end; *k;
end; *i;
psik[s,] = shape(pik,1,&bttotal);
end; *s;

ptheta = j(&n1,&cumbtotal,0);

do s = 1 to &n1;
do i = 1 to &i2;
do j = 1 to 5;
temp2 = (i-1)*5+j;
temp3 = (i-1)*4+j;

*discrete item scores are computed below;

if (j = 1) then ptheta[s,temp2] = 1 - psik[s,temp3];
else if (j >= 2 & j <= 4) then ptheta[s,temp2] = psik[s,temp3-1]
- psik[s,temp3];
else if (j = 5) then ptheta[s,temp2] = psik[s,temp3-1];
end;
end;
end;

use &polyitemscores_rg;      *using the item scores for rg
generated from mplus;
read all var ('i1':'i10') into itemscores;
close &polyitemscores_rg;

*essi for each subject (sum of psik for each item) is computed
below;
*also the total score for each subject is computed below;

itemscores = itemscores + 1;

ESRrg = j(&n1,&i2,0);
TRrg = j(&n1,1,0);
do s = 1 to &n1;
do i = 1 to &i2;
ESRrg[s,i] = ptheta[s,(i-1)*5+(itemscores[s,i])]*itemscores[s,i];
end;

```

```

TRrg[s] = sum(ESRrg[s,1:&i2]);
end;

create ES_Ref_rg from ESRrg[colname = {i1 i2 i3 i4 i5 i6 i7 i8 i9
i10}];
append from ESRrg;
close ES_Ref_rg;

Truescore = {TRefRG};
create T_Ref_rg from TRrg[COLNAME=Truescore];
append from TRrg;
close T_Ref_rg;

quit;

proc iml;
use &fg_poly;
*the fg item parameters for the 10 non-dif polytomous items;
read all var ('A1':'A2') into afg;
read all var ('B1':'B4') into bfg;
close &fg_poly;

use fscores_rg;      *the theta estimates for rg from mplus;
read all var {f1 f2} into fscores;
close fscores_rg;

psik = j(&n1,&bttotal,0);

do s = 1 to &n1;
pik = j(&i2,4,0);
do i = 1 to &i2;
do k = 1 to 4;
temp1 = 0;
do h = 1 to 2;
temp1 = (temp1 + afg[i,h]*(fscores[s,h] - bfg[i,k]));
end; *h;
pik[i,k] = exp(temp1)/(1+exp(temp1));
end; *k;
end; *i;
psik[s,] = shape(pik,1,&bttotal);
end; *s;

ptheta = j(&n1,&cumbtotal,0);

do s = 1 to &n1;
do i = 1 to &i2;
do j = 1 to 5;
temp2 = (i-1)*5+j;
temp3 = (i-1)*4+j;

*discrete item scores are computed below;

```

```

if (j = 1) then ptheta[s,temp2] = 1 - psik[s,temp3];
else if (j >= 2 & j <= 4) then ptheta[s,temp2] = psik[s,temp3-1]
- psik[s,temp3];
else if (j = 5) then ptheta[s,temp2] = psik[s,temp3-1];

end;
end;
end;

use polyitemscores_rg;
*using the item scores for rg generated from mplus;
read all var ('i1':'i10') into itemscores;
close polyitemscores_rg;

*ESSi for each subject (sum of psik for each item) is computed
below;
*also the total score for each subject is computed below;

itemscores = itemscores + 1;

ESRfg = j(&n1,&i2,0);
TRfg = j(&n1,1,0);
do s = 1 to &n1;
do i = 1 to &i2;
ESRfg[s,i] = ptheta[s,(i-1)*5+(itemscores[s,i])]*itemscores[s,i];
end;
TRfg[s] = sum(ESRfg[s,1:&i2]);
end;

create ES_Ref_fg from ESRfg[colname = {i1 i2 i3 i4 i5 i6 i7 i8 i9
i10}];
append from ESRfg;
close ES_Ref_fg;

Truescore = {TRefFG};
create T_Ref_fg from TRfg[COLNAME=Truescore];
append from TRfg;
close T_Ref_fg;

quit;

*using total test scores for rg and fg to compute dtf;

data t_ref_rg;
set T_Ref_rg;
Subject = _N_;
run;

data T_Ref_fg;
set T_Ref_fg;
Subject = _N_;
run;

```

```

data T_Ref;
merge T_Ref_rg T_Ref_fg;
by Subject;
drop subject;
run;

data D;
set T_Ref;
D = (TRefRG-TRefFG)**2;
*D = DsquareS**.5;
run;

proc means noprint data = D;
var TRefRG TRefFG D;
output out =
MeanD MEAN(TRefRG TRefFG D) = MeanTRefRG MeanTRefFG MeanD
STDDEV(TRefRG TRefFG D) = StddevTRefRG StddevTRefFG StddevD;
run;

data MeanD;
set MeanD;
keep MeanTRefRG MeanTRefFG MeanD StddevTRefRG StddevTRefFG
StddevD;
run;

proc iml;
use MeanD;
read all var {MeanTRefRG MeanTRefFG MeanD StddevTRefRG
StddevTRefFG StddevD} into MDTF;
close MeanD;

use Mean1;
read all var {StddevD} into VARD;
close Mean1;

DTF = j(1,1,0);
DTF = MDTF[,6]**2 + MDTF[,3]**2;

chil = DTF*&n1;
chisq = chil / (VARD[,1]**2);
p = 1 - (probchi(chisq,&n1));

chiDTF = DTF||chisq||p;

create chisqDTF1 from chiDTF[colname = {DTF chisq p}];
append from chiDTF;
close chisqDTF1;

quit;

data chisqDTF;
set chisqDTF1;

```

```
proc append BASE = chisqDTF DATA=chisqDTF1;  
run;  
  
%mend dfit;  
  
%dfit (n1=&n1, fg_poly=fg_poly_no_DIF, rg_poly=rg_poly_no_DIF,  
polyitemscores_rg=polyitemscores_rg_no_DIF, i2=10, bttotal=40,  
cumbtotal=50);
```


E.5 MGRM-DFIT FOR THE 10 NON-DIF ITEMS + ONE DIF ITEM ADDED

AT A TIME

```
/******  
This step performs the MGRM DFIT on the 10 polytomous items with  
no DIF. In addition, one DIF item at a time are added to the 10  
non-DIF items, therefore, DFIT is performed on 11 items at each  
instance.
```

```
THE FOLLOWING MACRO VARIABLES NEED TO BE SPECIFIED HERE:
```

```
- nl for group 1 (since RG item scores are used for both RG and  
FG computations of ESSi)
```

```
only the 'n' for the RG will be used for DFIT computations...
```

```
THE FOLLOWING MACRO VARIABLES WILL CHANGE FOR EACH DFIT TEST  
PERFORMED, BUT NOT CHANGE FROM RUN TO RUN IN THIS FILE:
```

```
- fg_poly_no_DIF (this includes the FG polytomous item parameters  
for the 10 non-DIF items, and one of the DIF items added)
```

```
- rg_poly_no_DIF (this includes the RG polytomous item parameters  
for the 10 non-DIF items, and one of the DIF items added)
```

```
- i2 - refers to the total number of items for this DFIT chisq  
testing, and it will refer to 11 items, at each instance here
```

```
- btotat - refers to the total number of 'b' parameters - for a  
5-point scale, this would refer to 4*11 = 44
```

```
- cumbtotat - refers to the total number of categories - for a 5-  
point scale, this would be 5*11 = 44
```

```
THE FOLLOWING SAS DATA FILES ARE CREATED IN THIS STEP:
```

```
- ES_Ref_rg - required for CDIF & NCDIF testing
```

```
- ES_Ref_fg - required for CDIF & NCDIF testing
```

```
- T_Ref_rg - required for DTF testing
```

```
- T_Ref_fg - required for DTF testing
```

```
- T_Ref - combined file of T_Ref_rg and T_Ref_fg used for  
computing DsquareS
```

```
- D - file need computing NCDIF and CDIF
```

```
- MeanD - needed for chisquare testing
```

```
- chisqDTF1 - a file that contains the chisq test for the  
particular DFIT test
```

```
- chisqDTF - a file that contains the appended chisq tests for  
all DFIT tests, this file will be appended to after each chisq  
test, in this and the following steps
```

```
*****/
```

```
%macro dfit (nl= , fg_poly= , rg_poly= , polyitemscores_rg= , i2=  
 , btotat= , cumbtotat=);
```

```
proc iml;
```

```
use &rg_poly;
```

```
*the rg item parameters for the 11 items tested in each case;
```

```
read all var ('A1':'A2') into arg;
```

```
read all var ('B1':'B4') into brg;
```

```

close &rg_poly;

use fscores_rg; *the theta estimates for rg from Mplus;
read all var {f1 f2} into fscores;
close fscores_rg;

psik = j(&n1,&bttotal,0);

do s = 1 to &n1;
pik = j(&i2,4,0);
do i = 1 to &i2;
do k = 1 to 4;
temp1 = 0;
do h = 1 to 2;
temp1 = (temp1 + arg[i,h]*(fscores[s,h] - brg[i,k]));
*the MGRM psik is computed using parms generated from Mplus
baseline;
end; *h;
pik[i,k] = exp(temp1)/(1+exp(temp1));
end; *k;
end; *i;
psik[s,] = shape(pik,1,&bttotal);
end; *s;

ptheta = j(&n1,&cumbtotal,0);

do s = 1 to &n1;
do i = 1 to &i2;
do j = 1 to 5;
temp2 = (i-1)*5+j;
temp3 = (i-1)*4+j;

*discrete item scores are computed below;

if (j = 1) then ptheta[s,temp2] = 1 - psik[s,temp3];
else if (j >= 2 & j <= 4) then ptheta[s,temp2] = psik[s,temp3-1]
- psik[s,temp3];
else if (j = 5) then ptheta[s,temp2] = psik[s,temp3-1];

end;
end;
end;

use &polyitemscores_rg;
*using the item scores for RG generated from Mplus;
read all var ('i1':'i11') into itemscores;
close &polyitemscores_rg;

*ESSi for each subject (sum of Psik for each item) is computed
below;
*also the total score for each subject is computed below;

```

```

itemscores = itemscores + 1;

ESRrg = j(&n1,&i2,0);
TRrg = j(&n1,1,0);
do s = 1 to &n1;
do i = 1 to &i2;
ESRrg[s,i] = ptheta[s,(i-1)*5+(itemscores[s,i])]*itemscores[s,i];
end;
TRrg[s] = sum(ESRrg[s,1:&i2]);
end;

create ES_Ref_rg from ESRrg[colname = {i1 i2 i3 i4 i5 i6 i7 i8 i9
i10 i11}];
append from ESRrg;
close ES_Ref_rg;

Truescore = {TRefRG};
create T_Ref_rg from TRrg[COLNAME=Truescore];
append from TRrg;
close T_Ref_rg;

quit;

proc iml;
use &fg_poly;
*the fg item parameters for the 11 items tested in each case;
read all var ('A1':'A2') into afg;
read all var ('B1':'B4') into bfg;
close &fg_poly;

use fscores_rg; *the theta estimates for rg from Mplus;
read all var {f1 f2} into fscores;
close fscores_rg;

psik = j(&n1,&btotol,0);

do s = 1 to &n1;
pik = j(&i2,4,0);
do i = 1 to &i2;
do k = 1 to 4;
temp1 = 0;
do h = 1 to 2;
temp1 = (temp1 + afg[i,h]*(fscores[s,h] - bfg[i,k]));
*the MGRM psik is computed using parms generated from Mplus
baseline;
end; *h;
pik[i,k] = exp(temp1)/(1+exp(temp1));
end; *k;
end; *i;
psik[s,] = shape(pik,1,&btotol);
end; *s;

```

```

ptheta = j(&n1,&cumbtotal,0);

do s = 1 to &n1;
do i = 1 to &i2;
do j = 1 to 5;
temp2 = (i-1)*5+j;
temp3 = (i-1)*4+j;

*discrete item scores are computed below;

if (j = 1) then ptheta[s,temp2] = 1 - psik[s,temp3];
else if (j >= 2 & j <= 4) then ptheta[s,temp2] = psik[s,temp3-1]
- psik[s,temp3];
else if (j = 5) then ptheta[s,temp2] = psik[s,temp3-1];

end;
end;
end;

use polyitemscores_rg;
*using the item scores for RG generated from Mplus;
read all var ('i1':'i11') into itemscores;
close polyitemscores_rg;

*ESSi for each subject (sum of Psik for each item) is computed
below;
*also the total score for each subject is computed below;

itemscores = itemscores + 1;

ESRfg = j(&n1,&i2,0);
TRfg = j(&n1,1,0);
do s = 1 to &n1;
do i = 1 to &i2;
ESRfg[s,i] = ptheta[s,(i-1)*5+(itemscores[s,i])]*itemscores[s,i];
end;
TRfg[s] = sum(ESRfg[s,1:&i2]);
end;

create ES_Ref_fg from ESRfg[colname = {i1 i2 i3 i4 i5 i6 i7 i8 i9
i10 i11}];
append from ESRfg;
close ES_Ref_fg;

Truescore = {TRefFG};
create T_Ref_fg from TRfg[COLNAME=Truescore];
append from TRfg;
close T_Ref_fg;

quit;

*using total test scores for rg and fg to compute dtf;

```

```

data T_Ref_rg;
set T_Ref_rg;
Subject = _N_;
run;

data T_Ref_fg;
set T_Ref_fg;
Subject = _N_;
run;

data T_Ref;
merge T_Ref_rg T_Ref_fg;
by Subject;
drop subject;
run;

data D;
set T_Ref;
D = (TRefRG-TRefFG)**2;
*D = DsquareS**.5;
run;

proc means noprint data = D;
var TRefRG TRefFG D;
output out =
MeanD MEAN(TRefRG TRefFG D) = MeanTRefRG MeanTRefFG MeanD
STDDEV(TRefRG TRefFG D) = StddevTRefRG StddevTRefFG StddevD;
run;

data MeanD;
set MeanD;
keep MeanTRefRG MeanTRefFG MeanD StddevTRefRG StddevTRefFG
StddevD;
run;

proc iml;
use MeanD;
read all var {MeanTRefRG MeanTRefFG MeanD StddevTRefRG
StddevTRefFG StddevD} into MDTF;
close MeanD;

use Mean1;
read all var {StddevD} into VARD;
close Mean1;

DTF = j(1,1,0);
DTF = MDTF[,6]**2 + MDTF[,3]**2;

chil = DTF*&n1;
chisq = chil / (VARD[,1]**2);
p = 1 - (probchi(chisq,&n1));

```

```

chiDTF = DTF||chisq||p;

create chisqDTF1 from chiDTF[colname = {DTF chisq p}];
append from chisqDTF;
close chisqDTF1;

quit;

data chisqDTF;
set chisqDTF;
proc append BASE = chisqDTF DATA=chisqDTF1;
run;

%mend dfit;

%dfit(n1=&n1,fg_poly=fg_poly_no_DIF_11,rg_poly=rg_poly_no_DIF_11,
polyitemscores_rg=polyitemscores_rg_no_DIF_11,i2=11,btotal=44,cum
btotal=55);
%dfit(n1=&n1,fg_poly=fg_poly_no_DIF_12,rg_poly=rg_poly_no_DIF_12,
polyitemscores_rg=polyitemscores_rg_no_DIF_12,i2=11,btotal=44,
cumbtotal=55);
%dfit(n1=&n1,fg_poly=fg_poly_no_DIF_13,rg_poly=rg_poly_no_DIF_13,
polyitemscores_rg=polyitemscores_rg_no_DIF_13,i2=11,btotal=44,
cumbtotal=55);
%dfit(n1=&n1,fg_poly=fg_poly_no_DIF_14,rg_poly=rg_poly_no_DIF_14,
polyitemscores_rg=polyitemscores_rg_no_DIF_14,i2=11,btotal=44,
cumbtotal=55);

```

E.6 SAVING THE MGRM-DFIT CHI-SQUARE TESTS FOR ALL COMPARATIVE MODELS

```
/******  
This is the final step required to create the chisq file for the  
DFIT chi-square tests for all models in each replication.
```

```
Once all the chi square tests are performed and appended to in  
the chisqDTF sas data file, this step creates the .dat file  
required to append across replications.
```

```
The following sas data sets and .dat files are created:
```

```
- chisqDTF
```

```
- chisqDTF.dat
```

```
*****/
```

```
data _null_;  
set chisqDTF;  
file 'c:\dif\chisqDTF.dat';  
put (_all_) ('09'X);  
run;
```

APPENDIX F

SAS CODE FOR COMPILING THE SIMULATION

F.1 COMPILING THE SIMULATION AND AUTOMATING ACROSS REPLICATIONS

```
libname project 'C:\DIF';  
%let nrep = 1;  
%let amag = 0.2;  
%let bmag = 0.5;
```

```
data means1;  
input m;  
datalines;  
0  
0  
;  
run;
```

```
data means2;  
input m;  
datalines;  
0  
-0.5  
;  
run;
```

```
data means3;  
input m;  
datalines;  
-0.5  
0  
;  
run;
```

```
data means4;  
input m;  
datalines;  
-0.5  
-0.5  
;  
run;
```



```

%let meansFG1=means1;
%let meansFG2=means2;
%let meansFG3=means3;
%let meansFG4=means4;

options mprint;

%macro loopstudy;

%do rep = 1 %to &nrep;

%do f1=1 %to 1;
/*the 1st factor will manipulate "sample size": 1000 & 2000*/
    %if &f1=1 %then %let n = 10000;
    %else %if &f1=2 %then %let n = 2000;
    %else %if &f1=3 %then %let n = 5000;

%do f2=1 %to 3;
/*the 2nd factor will manipulate the "SSR": 80/20, 70/30 and
50/50*/

    %if &f2=1 %then %do;
        %let nlratio = 0.80; %let n2ratio = 0.20; %end;
    %else %if &f2=2 %then %do;
        %let nlratio = 0.70; %let n2ratio = 0.30; %end;
    %else %if &f2=3 %then %do;
        %let nlratio = 0.50; %let n2ratio = 0.50; %end;
        %let n1 = &n * &nlratio;
        %let n2 = &n * &n2ratio;

%do f3=1 %to 4;
/*the 3rd factor will manipulate "uniform DIF" in the 'b'
parameter for DIF only in the last 'b' parameter, or DIF in the
last two 'b's with "DIF Direction": RG>FG or FG>RG*/

    %if &f3=1 %then %do;
        %let b1 = 0; %let b2 = 1; %end;
    %else %if &f3=2 %then %do;
        %let b1 = 1; %let b2 = 1; %end;
    %else %if &f3=3 %then %do;
        %let b1 = 0; %let b2 = -1; %end;
    %else %if &f3=4 %then %do;
        %let b1 = -1; %let b2 = -1; %end;
        %let bmag1 = &bmag * &b1;
        %let bmag2 = &bmag * &b2;

```

```

%do f4=1 %to 5;
/*the 4th factor will manipulate "non-uniform DIF" in the 'a'
parameter with "DIF in One Dimension": with DIF in either
"dimension 1" or "dimension 2" or "DIF in both dimensions" with
DIF "in the same direction" or "in opposite direction", "DIF
Direction": RG>FG or FG>RG*/
    %if &f4=1 %then %do;
        %let a1 = 0; %let a2 = 1; %end;
    %else %if &f4=2 %then %do;
        %let a1 = 1; %let a2 = 0; %end;
    %else %if &f4=3 %then %do;
        %let a1 = 1; %let a2 = 1; %end;
    %else %if &f4=4 %then %do;
        %let a1 = 1; %let a2 = -1; %end;
    %else %if &f4=5 %then %do;
        %let a1 = -1; %let a2 = 1; %end;
        %let amag1 = &amag * &a1;
        %let amag2 = &amag * &a2;

%do f5=1 %to 4;
/*the 5th factor will manipulate "mean differences": 0 0, 0 -.5,
-.5 0, -.5 -.5*/

    %if &f5=1 %then %let meansFG = &meansFG1;
    %else %if &f5=2 %then %let meansFG = &meansFG2;
    %else %if &f5=3 %then %let meansFG = &meansFG3;
    %else %if &f5=4 %then %let meansFG = &meansFG4;

%let cond = %eval
(1000*&f5+10000*&f4+100000*&f3+1000000*&f2+10000000*&f1);

%let seed=%eval(&cond + &rep);

* suppress output and log file;

proc printto log=out print=out new;
run;

```

```

x 'cd C:\DIF';

%include 'gen_bivariate_normal_data.sas';
%include 'MGRM_discrete_data_5point_IML.sas';
x "echo &f1 &f2 &f3 &f4 &f5 &rep >>
  randdich_a_start_all.dat";
x 'type randdich_a_start.dat >>
  randdich_a_start_all.dat';
x "echo &f1 &f2 &f3 &f4 &f5 &rep >>
  randpoly_a_start_all.dat";
x 'type randpoly_a_start.dat >>
  randpoly_a_start_all.dat';
x 'mplus dif_baseline_5.inp';
x 'mplus dif_measurementinvariance_51.inp';
x 'mplus dif_measurementinvariance_52.inp';
x 'mplus dif_measurementinvariance_53.inp';
x 'mplus dif_measurementinvariance_54.inp';
x 'mplus dif_measurementinvariance_55.inp';
x 'mplus dif_measurementinvariance_56.inp';
%include 'read_data_5_point.sas';
x "echo &f1 &f2 &f3 &f4 &f5 &rep >>
  chisq_mplus_all.dat";
x 'type chisq_mplus.dat >>
  chisq_mplus_all.dat';
%include 'read_factorscores_from_Mplus.sas';
%include 'read_chisq_diff_from_Mplus.sas';
x "echo &f1 &f2 &f3 &f4 &f5 &rep >>
  Chisq_diff_all.dat";
x 'type Chisq_diff.dat >>
  Chisq_diff_all.dat';
%include 'creating_datasets_for_DFIT_chisq_testing.sas';
%include 'DFIT_all_items.sas';
%include 'DFIT_CDIF_NCDIF.sas';
x "echo &f1 &f2 &f3 &f4 &f5 &rep >> CDIFi_all.dat";
x 'type CDIFi.dat >> CDIFi_all.dat';
x "echo &f1 &f2 &f3 &f4 &f5 &rep >> NCDIFi_all.dat";
x 'type NCDIFi.dat >> NCDIFi_all.dat';
%include 'DFIT_no_DIF_items.sas';
%include 'DFIT_no_DIF_items_adding_11-14.sas';
%include 'creating_chisqDTF_for_reps.sas';
x "echo &f1 &f2 &f3 &f4 &f5 &rep >> chisqDTF_all.dat";
x 'type chisqDTF.dat >> chisqDTF_all.dat';

%end; * f5;
%end; * f4;
%end; * f3;
%end; * f2;
%end; * f1;
%end; * loop rep;

%mend;

```

```
data;  
options noxwait;  
x "cd C:\DIF";  
%loopstudy  
run;  
  
proc printto;  
run;  
  
options mprint;
```

BIBLIOGRAPHY

- Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement*, 13, 113-127.
- Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1-23.
- Adams, R. J., Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1-23.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.
- Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, 9, 37-48.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, DC: American Council on Education.
- Angoff, W. H. (1972, Sept). *A technique for the identification of cultural differences*. Paper presented at the annual meeting of the American Psychological Association, Honolulu, HI.
- Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp.3-23). Hillside: Lawrence Erlbaum Associates.
- Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, 10, 95-106.

- Angoff, W. H., & Sharon, A. T. (1974). The evaluation of differences in test performance of two or more groups. *Educational and Psychological Measurement*, 34, 807-816.
- Bagozzi, R. P., & Edwards, J. R. (1998). A general approach for representing constructs in organizational research. *Organizational Research Methods*, 1, 45-87.
- Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement*, 16, 87-96.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.
- Bock, D. R., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-459.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.
- Bolt, D. M. (2002). A monte carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, 15, 113-141.
- Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105, 456-466.
- Camili, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement*, 16, 129-147.
- Camili, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioral Statistics*, 24, 323-341.
- Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, 12, 253-260.
- Cardall, C., & Coffman, W. E. (1964). A method for comparing the performance of different groups on the same items of a test. *Research and Development Reports*, 9, 64-65. Princeton, NJ: Educational Testing Service.
- Chan, D. (2000). Detection of differential item functioning on the Kirton adaption-innovation inventory using the multiple-group mean and covariance structure analysis. *Multivariate Behavioral Research*, 35, 2, 169-199.

- Chang, H. H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika*, 59, 391-404.
- Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, 33, 333-353.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233-255.
- Cohen, A. S., & Kim, S-H. (1993). *A comparison of equating methods under the graded response model*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.
- Cohen, A. S., & Kim, S-H. (1998). An investigation of linking methods under the graded response model. *Applied Psychological Measurement*, 22, 2, 116-130.
- Cohen, A. S., Kim, S-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement*, 17, 335-350.
- Cole, N. S. (1993). History and development of differential item functioning. In P. W. Holland & H. Wainer (Eds.). *Differential item functioning* (pp.25-33). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Collins, W. C., Raju, N. S., & Edwards, J. E. (2000). Assessing differential functioning in a satisfaction scale. *Journal of Applied Psychology*, 85, 451-461.
- Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques. *Medical Care*, 44, 11, S115-S123.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace.
- Davey, T., Oshima, T. C., & Lee, K. (1996). Linking multidimensional item calibrations. *Applied Psychological Measurement*, 20, 4, 405-416.
- De Ayala, R. J. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement*, 18, 2, 115-170.
- De la Torre, J., & Patz, R. J. (2002, April). *A multidimensional item response theory approach to simultaneous ability estimation*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

- De la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, 30, 295-311.
- Drasgow, F., & Knafer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology*, 70, 662-680.
- Edelen, M. O., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, J. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach. *Medical Care*, 44, 11, S134-S142.
- Edwards, M. C (2005). A markov chain monte carlo approach to confirmatory item factor analysis. Unpublished doctoral dissertation, University of North Carolina – Chapel Hill.
- Edwards, M. C., & Wirth, R. J. (2009). Measurement and the study of change. *Research in Human Development*, 6, 2-3, 74-96.
- Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Fan, X., & Sivo, S. A. (2009). Using Δ goodness-of-fit indexes in assessing mean structure invariance. *Structural Equation Modeling*, 16, 54-69.
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9 (4), 466-491.
- Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement*, 23, 309-326.
- Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement*, 13, 373-389.
- Glockner-Rist, A., & Hoijsnik, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling*, 10, 4, 544-565.
- Gonzales-Roma, V., Hernandez, A., & Gomez-Benito, J. (2006). Power and Type-I error of the mean and covariance structure analysis model for detecting differential item functioning in graded response items. *Multivariate Behavioral Research*, 41, 1, 29-53.

- Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21, 347-360.
- Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, 22, 144-149.
- Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care*, 44, 11, S182-S188.
- Harwell, M, Stone, C.A., Hsu, T.C., & Kirisci, L. (1996) Monte Carlo Studies in Item Response Theory. *Applied Psychological Measurement*, 20, 101-126.
- Hill, C. D., Edwards, M. C., Thissen, D., Langer, M. M., Wirth, R. J., Burwinkle, T. M., & Varni, J. W. (2007). Practical issues in the application of item response theory: A demonstration using items from the Pediatric quality of life inventory (PedsQL) 4.0 generic core scales. *Medical Care*, 45, 5, S39-S47.
- Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations*. Thousand Oaks, CA: Sage.
- Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer, & H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Holland, P. W., & Wainer, H. (1993). Introduction and background to differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. xiii-xv; 1-3). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, 18, 117-144.
- Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo Study. *Applied Psychological Measurement*, 6 (3), 249-260.
- Jones, R. N. (2006). Identification of measurement differences between english and spanish language versions of the mini-mental state examination: Detecting differential item functioning using MIMIC modeling. *Medical Care*, 44, 11, S124-S133.
- Jöreskog, K. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 36, 409-426.

- Jöreskog, K., & Sörbom, D. (1996). *LISREL 8 user's guide*. Chicago: Scientific software.
- Kannan, P., & Kim, K. H. (2009, April). *Item parameter recovery for a within-item multidimensional graded response model: A SEM-CFA perspective*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.
- Kannan, P., & Ye, F. (2008, April). *Item parameter recovery for a between-item multidimensional graded response model*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.
- Kim, S-H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement*, 22, 345-355.
- Kim, S-H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, 26, 1, 25-41.
- Kirisci, L., Hsu, T-C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement*, 25 (2), 146-162.
- Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices*. 2nd ed. New York, NY: Springer-Verlag.
- Koretz, D., & Hamilton, L. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational Measurement* (4th edition), (pp. 531-542). Westport, CT: American Council on Education/Praeger.
- Kyllonen, P.C. (September, 2008). Enhancing noncognitive skills to boost academic achievement. In *Educational Testing in America: State Assessments, Achievement Gaps, National Policy and Innovations (Session III: Innovations in Testing)*. Willard Hotel, Washington, DC.
- Kyllonen, P. C., Walters, A. M., & Kaufman, J. C. (2005). Noncognitive constructs and their assessment in graduate education: A review. *Educational Assessment*, 10, 3, 153-184.
- Lane, S., Wang, N., & Magone, M. (1996). Gender-related differential item functioning on a middle school mathematics performance assessment. *Educational Measurement: Issues and Practice*, 15, 4, 21-27, 31.
- Lane, S., & Stone, C. A. (2006). Performance Assessment. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.). Westport, CT: American Council on Education.

- Lee, G., Kolen, M. J., Frisbie, D. A., & Ankemann, R. D. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Psychological Measurement*, 25, 4, 357-372.
- Lee, K., & Oshima, T. C., (1996). Multidimensional and unidimensional item parameter linking in item response theory. *Applied Psychological Measurement*, 20, 3, pp. 230.
- Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement*, 5, 159-173.
- Little, T. D. (1997). Mean and covariance structure (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, 32, 53-76.
- Lloyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, 17, 179-193.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Macro, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 14, 139-160.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- Maurer, T. J., Raju, N. S., & Collins, W. C. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology*, 83, 5, 693-702.
- Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1998). Multidimensional DIF analyses: The effect of matching on unidimensional subtest scores. *Applied Psychological Measurement*, 22, 4, 357-367.
- McKinley, R. L., & Reckase, M. D. (1983, April). *The use of IRT analysis on dichotomous data from multidimensional tests*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec.
- Meade, A. W., & Lautenschlager, G. J. (2004a). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, 7, 4, 361-388.
- Meade, A. W., & Lautenschlager, G. J. (2004b). A monte-carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling*, 11, 1, 60-72.

- Meade, A. W., Lautenschlager, G. J., & Johnson, E. C. (2006, April). *Alternative cutoff values and DFIT tests of measurement invariance*. Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525-543.
- McLeod, L. D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test Scoring* (pp. 187-216). Mahwah, NJ: Lawrence Erlbaum.
- Miller, D. M., & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement*, 16, 381-388.
- Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research*, 30, 577-605.
- Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, 17, 297-334.
- Morales, L. S., Flowers, C., Guttierrez, P., Kleinman, M., & Teresi, J. A. (2006). Item and scale differential functioning of the mini-mental state exam assessed using the differential item and test functioning (DFIT) framework. *Medical Care*, 44, 11, S143-S151.
- Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, 14, 59-71.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, 19 (1), 73-90.
- Muthén, B. O., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus: Mplus Web Note #4 (www.statmodel.com). (Version 5.0). Los Angeles, CA: Muthén & Muthén.
- Muthén, L. K., & Muthén, B. O. (2006a). Mplus: Statistical analysis with latent variables (Version 5.1). Los Angeles, CA: Muthén & Muthén.
- Muthén, L. K., & Muthén, B. O. (2006b). Mplus Technical Appendix: IRT in Mplus (Version 5.1). Los Angeles, CA: Muthén & Muthén.

- Oshima, T. C., & Morris, S. B. (2008). An NCME instructional module on Raju's differential functioning of items and tests (DFIT). *Instructional Topics in Educational Measurements (ITEMS): The National Council of Measurement in Education, Fall 2008*, 43-50.
- Oshima, T. C., Davey, T. C., & Lee, K. (2000). Multidimensional linking: Four practical approaches. *Journal of Educational Measurement*, 37, 4, 357-373.
- Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests. *Journal of Educational Measurement*, 34, 253-272.
- Potenza, M. T., & Dorens, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement*, 19, 23-37.
- Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, 14, 197-207.
- Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, 87, 517-529.
- Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, 19, 353-368.
- Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9, 401-412.
- Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21, 25-36.
- Reckase, M. D., & McKinley, R. L. (1991). The discrimination power of items that measure more than one dimension. *Applied Psychological Measurement*, 15, 361-373.
- Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, 114, 552-566.
- Roberts, J. (2009, April). Bringing different perspectives to a reflect a common issue: Discussant Comments. In L. Yao (Chair). *Estimation Issues in Multidimensional IRT*. Paper session conducted at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

- Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, 20, 4, 355-371.
- Rubin, D. B. (1988). Discussion: On Holland and extending the DIF model. In H. Wainer & H. L. Braun (Eds.), *Test validity* (pp. 246-247). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.
- Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph*, No. 18.
- Schmitt, N. (1982). The use of analysis of covariance structures to assess beta and gamma change. *Multivariate Behavioral Research*, 17, 343-358.
- Schulz, W. (2005). *Testing parameter invariance for questionnaire indices using confirmatory factor analysis and item response theory*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco: CA.
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, 91, 6, 1292-1306.
- Stocking, M., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 207-210.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361-370.
- Swygert, K. A., McLeod, L. D., & Thissen, D. (2001). Factor analysis for items or testlets scored in more than two categories. In D. Thissen & H. Wainer (Eds.), *Test Scoring* (pp. 217-250). Mahwah, NJ: Lawrence Erlbaum.
- Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52 (3), 393-408.
- Tate, R. L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement*, 36, 4, 36-346.
- Teresi, J. A. (2006a). Overview of quantitative measurement methods: Equivalence, invariance, and differential item functioning in health applications. *Medical Care*, 44, 11, S39-S49.

- Teresi, J. A. (2006b). Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. *Medical Care*, 44, 11, S152-S170.
- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51, 4, 567-577.
- Thissen, D., Steinberg, L., & Gerard, M. (1986). Beyond mean group differences: The concept of item bias. *Psychological Bulletin*, 99, 118-128.
- Thomas, L. L., Kuncel, N. R., & Crede, M. (2007). Non-cognitive variables in college admissions: The case of the non-cognitive questionnaire. *Educational and Psychological Measurement*, 67, 4, 635-657.
- Vandenberg, R. J. & Lance, C. E. (2000). A review and synthesis of measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3, 4-70.
- Walker, C. M., & Beretvas, N. S. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement*, 38, 147-163.
- Wang, W-C. (2004). Effect of anchor item methods on the detection of differential item functioning within the family of rasch models. *The Journal of Experimental Education*, 72, 3, 221-261.
- Wang, W-C., Chen, P-H., & Cheng, Y-Y. (2004). Improving the measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, 9, 116-136.
- Wang, W.-C., Wilson, M. R., & Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. In M. Wilson, G. Engelhard, & K. Draney (Eds.), *Objective measurement: Theory into practice* (Vol. 4, pp. 139-155). Norwood, NJ: Ablex.
- Wang, W-C., & Yeh, Y-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, 27, 6, 479-498.
- Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and non-compensatory two-dimensional data on unidimensional IRT estimation. *Applied Psychological Measurement*, 12, 239-252.

- Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12 (1), 58-79.
- Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, 33, 1, 42-57.
- Wu, Q., & Lei, P-W. (2009, April). *Using multi-group confirmatory factor analysis to detect differential item functioning when tests are multidimensional*. Paper presented at the Annual Meeting of the National Council for Measurement in Education, San Diego: CA.
- Yang, F. M., Tommet, D., & Jones, R. N. (2009). Disparities in self-reported geriatric depressive symptoms due to sociodemographic differences: An extension of the bi-factor item response theory model for use in differential item functioning. *Journal of Psychiatric Research*, 43, 1025-1035.
- Yasuda, T., Lei, P-W., & Suen, H. K. (2007). Detecting differential item functioning in the Japanese version of the multiple affect adjective check list-revised. *Journal of Psychoeducational Assessment*, 25, 4, 373-384.
- Yao, L. (2004b). LinkMIRT: Linking of multivariate item response models [Computer software]. Monterey, CA: CTB/McGraw-Hill.
- Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, 30, 6, 469-492.
- Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, 31, 2, 83-105.
- Yao, L., & Boughton, K. A. (2009). Multidimensional linking for tests with mixed item types. *Journal of Educational Measurement*, 46, 2, 177-197.
- Zumbo, B. D. (2007). Three generations of DIF analysis: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4, 2, 223-233.